Noname manuscript No.
(will be inserted by the editor)

Intrackability: Characterizing Video Statistics and Pursuing Video Representations

Haifeng Gong · Song-Chun Zhu

the date of receipt and acceptance should be inserted later

Abstract Videos of natural environments contain a wide variety of motion patterns of varying complexities which are represented by many different models in the vision literature. In many situations, a tracking algorithm is formulated as maximizing a posterior probability. In this paper, we propose to measure the video complexity by the entropy of the posterior probability, called the intrackability, to characterize the video statistics and pursue optimal video representations. Based on the definition of intrackability, our study is aimed at three objectives. Firstly, we characterize video clips of natural scenes by intrackability. We calculate the intrackabilities of image points to measure the local inferential uncertainty, and collect the histogram of the intrackabilities over the video in space and time as the global video statistics. We find that a PCA scatter-plot based on the first two principle components of intrackability histograms can reflect the major variations, i.e., image scaling and object density, in natural video clips. Secondly, we show that different video representations, including deformable contours, tracking kernels with various appearance features, dense motion fields, and dynamic texture models, are connected by the change of intrackability and thus develop a simple criterion for model transition and for pursuing the optimal video representation. Thirdly, we derive the connections between the intrackability measure and other criteria in the literature such as the Shi-Tomasi texturedness measure, conditional number, and Harris-Stephens R score, and compare with the Shi-Tomasi measure in tracking experiments.

1 Introduction

1.1 Motivation and objective

Videos of natural environments contain a wide variety of motion patterns of varying complexities which are represented by many distinct models in the vision literature. Fig. 1 illustrates four typical representations: (i) A moving contour representing a slowly walking human figure in near view; (ii) A kernel (window with interior feature points) representing a fast moving car in middle distance; (iii) A dense motion (optical) flow field representing a marathon crowd motion; and (iv) An appearance based spatio-temporal autoregression (STAR) model representing the fire flame where it is hard to track any distinct elements. The complexity of these video clips are affected by a few major factors, namely, the object scale, the object density, and the stochasticity of the motion. Apparently, the change of these factors triggers transitions among these representations. Fig. 2 shows two sequences of motion at distinct scales: the bird flock and the marathon crowd, where the individual bird or person is represented by a contour, a kernel and a motion vector at three scales respectively.

These representations have been studied extensively for various tasks in the vision literature, for example, contour tracking (Maccormick and Blake, 2000; Sato and Aggarwal, 2004; Black and Fleet, 2000), kernel tracking (Comaniciu et al, 2003; Collins, 2003), PCA basis tracking (Ross et al, 2008; Kwon et al, 2009), motion vectors of points -– sparse (Shi and Tomasi, 1994; Tommasini et al, 1998; Segvič et al, 2006; Serby et al, 2004; Veenman et al, 2001) or dense (Horn and Schunck, 1981; Ali and Shah, 2007),

Haifeng Gong was a postdoc researcher in the Department of Statistics, UCLA (2007-2009) and a researcher and team leader at Lotus Hill Research Institute, China during (2006-2010), and is a postdoc researcher at Computer and Information Science in University of Pennsylvania since 2009. This work was done when he was in Lotus Hill Research Institute. Email address: hfgong@seas.upenn.edu.
Song-Chun Zhu is a professor at the department of statistics and department of computer science, UCLA and a founder of the Lotus Hill Institute, China. Email address: sczhu@stat.ucla.edu.

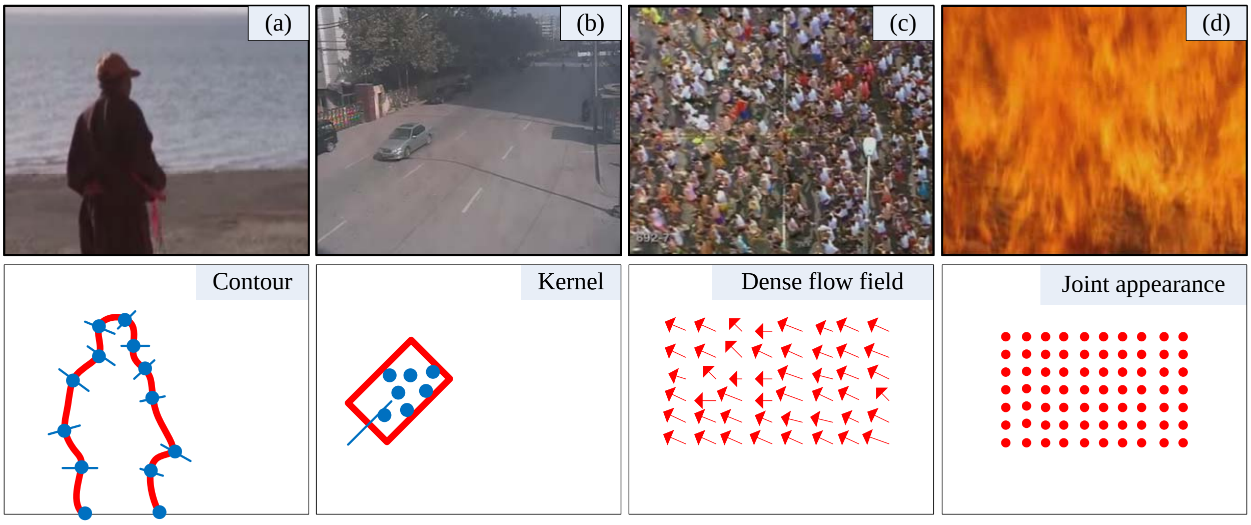

Fig. 1 Examples of motion patterns and their representations: (a) A slowly walking human figure at near view is represented by a contour; (b) A fast moving car in middle distance is represented by a kernel (window with multiple interior feature points); (c) A moving crowd in far view is represented by a dense motion field; and (d) the dynamic texture of fire has no distinct element that is trackable, and is represented by auto-regression models on its image intensities without explicit motion correspondence.

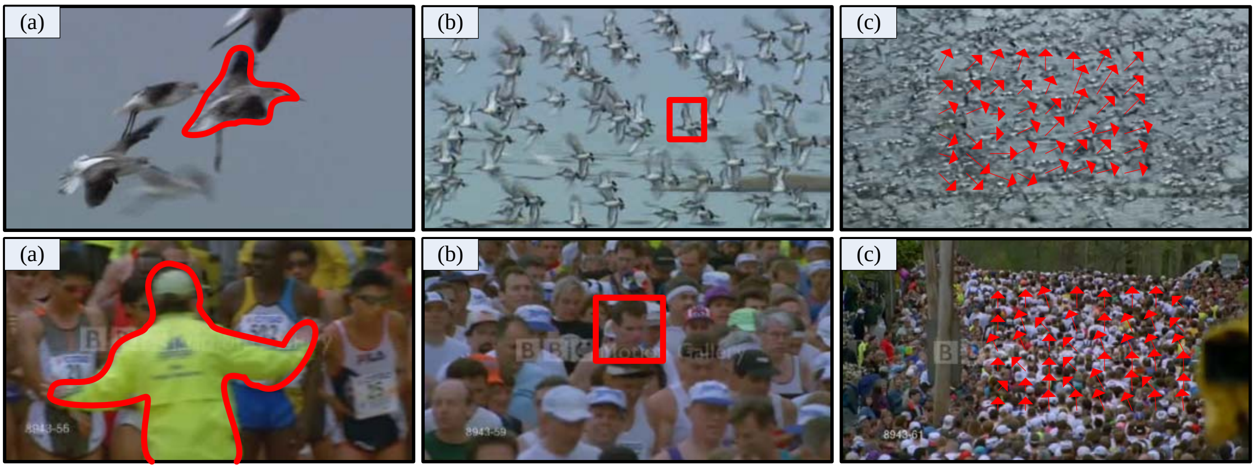

Fig. 2 The switch of video representations is triggered by image scaling (camera zooming) and density changes. (a) In high resolution, the bird shape and human figure are described by their contours; (b) In middle resolution, they are represented by a kernel with feature points; and (c) In low resolution, the people and birds are modeled by moving points with dense optical flow.

and dynamic texture (Szummer and Picard, 1996; Fitzgibbon, 2001; Soatto et al, 2001) or textured motion (Wang and Zhu, 2003). However, no attempt, to our best knowledge, has been made to formally characterize the video complexity and to establish connections and conditions for the transitions among these representations in the literature. In fact, the automated selection and switching of representations on-the-fly is of practical importance in real-time applications. For example, tracking an object over a long range of scales will need different representations. A surveillance system must also adapt its tracking task when the number of targets in a scene suddenly increases and cannot be tracked individually due to limitated computing resources. If the computing resource allows, it should output more detailed information for further processing, database indexing or human inspection. When the number of objects at near distance increases, heavy occlusions always happen and we have to change to track parts and discard some objects. When the number of objects at far distance increases, we can change to model motion flow and count number of objects. For example, (Ali and Shah, 2008) and (Cong et al, 2009) track high density crowd scenes with a motion field.

In this paper, we study an information theoretical criterion called the *intrackability* as a measure of the video complexity. By definition, the intrackability is the entropy of the posterior probability which a tracking or motion analysis algorithm tries to maximize, and thus reflects the difficulty and uncertainty in tracking certain elements (pixels, feature points, lines, patches). We will use the intrackability to characterize the video statistics, explain the transition between representations, and pursue the optimal video representation for an given video clip. (Here, to pursue means solving selection of image elements in a sequential greedy way.) More specifically, our study is aimed at the following three objectives.

Firstly, we are interested in characterizing the global statistics of video clips and developing a panoramic map for the variety of video representations. We calculate the intrackabilities of some atomic image elements (patches) to measure the local inferential uncertainty, and then we collect histograms of the intrackabilities over the video in space and time as the global video statistics. We find that these histograms can be roughly decomposed into three bands which correspond to three distinct motion regimes: (i) Low intrackability band for the trackable regime, which corresponds to image areas with distinct feature points or structured texture areas that can be tracked with high accuracy. (ii) High intrackability band for the intrackable regime, which corresponds to image areas with no distinct texture, for example, flat areas or extremely populous areas. (iii) Medium intrackability band which contains mostly texture areas where structures become less distinguishable. We use regimes to refer to them because they do not have a rigorous predefined boundaries like classes. Using a PCA analysis on these square root histograms, we find that the first two eigen-vectors represent two major changes in the video space: the transition between the trackable and the intrackable motion and the transition between structure and texture. We plot the scatter plot and map natural video clips to these two axes to gain insight into the variations of video complexity.

Secondly, we are interested in developing an information theoretical criterion to guide the transition and selection of video representations, in contrast to the common practice that the video representations are manually selected for different tasks. Our criterion is a sum of the intrackability of the tracked representation (W as a vector) and its complexity (the number of variables in W). By minimizing this criterion (over W), our algorithm automatically chooses an optimal representation for the video clip which is often hybrid – mixing various representations for different areas in the video. In the spectrum of representations, the most complex one is the dense motion flow where each pixel or feature point is tracked and W is a long vector, and the simplest one is the dynamic texture or textured motion where no velocity is computed as there are no distinct and trackable elements and W is a short statistical description of the motion impression. Intuitively, when the ambiguity (or intrackability) is large, we reduce the representation W by two ways: (i) dropping certain elements, for example, remove elements that are not trackable, or drop the motion direction in the tangent direction of a contour element; or (ii) merging some descriptions, for example, combining a number of feature points that have similar motion in a kernel. In experiments, we show that different video representations, including deformable contours, tracking kernels with various appearance features, dense motion fields, and spatial temporal auto-regression models are selected by the algorithm for different video clips.

Thirdly, we compare our intrackability measure with three other criteria in the literature: (i) the texturedness measure for good features to track (Shi and Tomasi, 1994), (ii) Harris R score (Harris and Stephens, 1988) for corner detection and (iii) the conditional number for robust tracking in (Fan et al, 2006). We show that all three measures are related to different formula of the two eigenvalues in the local Gaussian distribution over the possible velocity. The intrackability is a general measure that is closely related to the three criteria. We also compare the intrackability with Shi-Tomasi measure by tracking experiments.

1.2 Related work in the literature

For the first objective of characterizing statistics of video clips, our work is closely related to natural image statistics. For natural images, some interesting properties are observed in their histograms of filtered responses, such as high kurtosis, that leads to sparse coding and scale invariance in gradient histograms (we refer to (Srivastava et al, 2003) for a comprehensive review), and various image models are learned to account for these statistical observations. The work that most directly inspired our study is (Wu et al, 2008). In (Wu et al, 2008) the entropy of posterior probability is defined as *imperceptibility*, which is then shown theoretically to guide the transitions of our perception of images over object scales. In general, (Wu et al, 2008) identified three regimes of models along the axis of imperceptibility: (i) the low entropy regime for structured images (represented by sparse coding), (ii) the high entropy regime for textured images (represented by Markov random fields); and (iii) the Gaussian noise regime for flat images or images with stochastic texture. A perceptual scale space representation was studied in (Wang and Zhu, 2008). While these works characterize the statistical properties of image appearance, our study is focused on the global statistics of local motion. We replace the histograms of filtered responses by the histograms of local intrackability, which divide videos into various regimes of representations.

Our work is closely related to another stream of research — image scale-space theory, which was proposed by (Witkin, 1983) and (Koenderink, 1984) and extended by (Lindeberg, 1993). The Gaussian and Laplacian pyramids are two multiscale representations concerned in scale-space theory. A Gaussian pyramid is a series of low-pass filtered and down-sampled images. A Laplacian pyramid consists of band-passed images which are the difference between every two consecutive images in the Gaussian pyramid. Scale-space theory studied discrete and qualitative events, such as appearance of extremal points (Witkin, 1983), and tracking inflection points. The image scale-space theory has been widely used in vision tasks. In this paper, we study the motions of points, contours and kernels, rather than the appearances of image patches in terms of Gaussian and Laplacian pyramids. We study the transitions of these higher level representations over scales and object density, rather than appearance of extremal points and drifting of inflection points.

For the second and third objectives of representation pursuit and tracking feature selection, our work is related to the various criteria for feature selection (Marr et al, 1979; Dreschler and Nagel, 1981; Yilmaz et al, 2006) in the vast literature of motion analysis and tracking. The corner detector (Harris and Stephens, 1988) has been used as a tracking feature selector for years. It is defined on eigenvalues of a matrix collected from image gradients. For tracking based on sum-of-squared-differences (SSD), (Shi and Tomasi, 1994) selected good features by a texturedness measure which is also defined on the same matrix as (Harris and Stephens, 1988). (Nickels and Hutchinson, 2002) analyzed variations of probability distributions of SSD motion vectors, and measured the uncertainty in terms of a covariance matrix from Gaussian fitting. For tracking based on kernels, (Fan et al, 2006) gave a reliability measure for kernel features based on condition number of a linear equation system. Covariance is also used in (Zhou et al, 2005) as an uncertainty measure for SSD, MeanShift and shape matching. For multi-frame adaptive tracking, (Collins et al, 2005) used log likelihood ratio scores of objects against the background as a goodness measure. These measures are all associated with specified feature descriptions (e.g., SSD, kernel) and tracking model. A recent work (Pan et al, 2009) used a forward-backward tracking strategy to evaluate the robustness of a tracker — first the object is tracked forward for a few frames, then tracked backward from the end frame of forward tracking to the beginning one, and the difference of the initial position and the backward tracked result is used as a measure of the robustness.

There are numerous works in psychophysics, e.g. (Pylyshyn and Vidal Annan, 2006), that studied the human perception of motion uncertainty, and showed that human vision loses track of objects (dots) when the number of dots increases or their motion is too stochastic.

(Han et al, 2005) first proposed to use entropy to select the best template for tracking, but no detailed investigation was made. The authors proposed the intrackability concept in two short papers (Li et al, 2007b,a) in the context of surveillance tracking. The intrackability concept was also mentioned in (Badrinarayanan et al, 2007). The contents presented in this paper are much more general than these papers and are not published elsewhere. Another interesting work related to ours is (Kadir and Brady, 2001). They investigated the use of entropy measures to identify regions of saliency in scale space, and obtained reasonable results on a broad class of images and image sequences. They also used it for tracking feature selection. The key difference between their work and ours is that they use the entropy of image pixels, that is, they first collect the histogram from a template, then compute the entropy of the histogram; while we use the entropy of posterior probability of motion perception.

1.3 Contributions and paper plan

In summary, this paper makes the following contributions to the literature.

1. The paper defines intrackability quantitatively to measure inferential uncertainty and uses it to characterize video into different regimes of representations. Thus we draw some connections between different families of models in the motion/tracking literature.
2. The paper shows that intrackability can be used to pursue a hybrid representation composed of feature points, contours and kernels for various videos.
3. The paper shows that intrackability is a general criterion, and derives its relation to three other measures in the literature.

This paper is organized as follows. We first define intrackability and give a simple method for computing it on a simple probability model in Section 2. Then, we use the histogram of the intrackability measure to characterize natural videos in Section 3 and show the connections and transitions of different representations through scaling. Then in Section 4, we adopt the intrackability criterion for pursuing optimal video representations. Section 4 explains the relationship between intrackabilities and video representation. First, we give brief introductions of popular representations for motions in the literature. Then representation projection is introduced to explain how these representations can be convertible in a coarse-to-fine manner. Finally, based on a criterion considering both intrackability and level of details, an algorithm for automatic construction of hybrid representations is proposed, which produces representations that consist of feature points, contours and kernels. In Section 5, we show how intrackability is related to other criteria for selecting features to track. The paper is concluded in Section 6 with a discussion.

2 Intrackability: definition and computation

2.1 Definitions of intrackability

Let $\mathbf{I}(t)$ be an image defined on a window Λ at time t, and $\mathbf{I}[\tau] = (\mathbf{I}(1), \cdots, \mathbf{I}(\tau))$ a video clip in a time interval $[1, \tau]$, and W the representation of this video selected for various tasks, e.g., motion vectors, or positions of control points of contours. In a Bayesian view, the objective of motion analysis is to compute W by maximizing a posteriori probability

$$W^* = \arg\max_W p(W|\mathbf{I}[\tau]). \quad (1)$$

The optimal solution W^*, however, does not contain information about the uncertainty of the inference and can not tell whether the selected representation is appropriate for the video sequence. A common measure for the uncertainty is the entropy of the posterior probability, we call it the intrackability.

Definition 1 (video intrackability) Intrackability of a video sequence $\mathbf{I}_\Lambda[\tau]$ for a representation W is defined by,

$$\mathcal{H}\{W|\mathbf{I}[\tau]\} = -\sum_W p(W|\mathbf{I}[\tau]) \log p(W|\mathbf{I}[\tau]). \quad (2)$$

Here log is natural logarithm. We use the natural logarithm because it is more amenable to probability models of exponential family.

In this paper, we will focus on middle level representations that are local in space and time, e.g. pixels, points, lines, kernels etc., and W does not contain high level concepts, such as action and events. Thus the volume $\Lambda \times \tau$ is quite small. In a simplest case, $W = \mathbf{u}$ is the motion vector of a feature point, patch, or kernel and $\mathbf{I}$ and $\mathbf{I}'$ are two consecutive frames. Then the intrackability is $\mathcal{H}\{\mathbf{u}|\mathbf{I}, \mathbf{I}'\}$.

Definition 2 (local intrackability) Intrackability of a local element between two image frames $\mathbf{I}, \mathbf{I}'$ for its velocity $\mathbf{u}$ is $\mathcal{H}\{\mathbf{u}|\mathbf{I}, \mathbf{I}'\}$.

In the next two sections, we will use $\mathcal{H}\{\mathbf{u}|\mathbf{I}, \mathbf{I}'\}$ as a local intrackability to characterize the global video complexity.

In general, good features to track should be discriminative in both appearance and dynamics. Both factors are integrated in the intrackability measure, because the posterior probability $p(W|\mathbf{I}[\tau])$ encodes both appearance and motion information.

It is worth noting that $\mathcal{H}$ is an unbounded differential entropy for continuous variables W and $\mathbf{I}$. In this paper, we discretize both W and $\mathbf{I}$ in a finite set of values to obtain a non-negative bounded Shannon entropy.

2.2 Computing the local intrackability

The local intrackability can be exactly computed for a specified appearance and motion probability model. We take SSD appearance model with uniform motion prior as an example, in which the posterior probability is

$$p(\mathbf{u}|\mathbf{I}, \mathbf{I}') \propto \exp\left\{-\frac{\sum_{\mathbf{x}\in P} \|\mathbf{I}(\mathbf{x}) - \mathbf{I}'(\mathbf{x}+\mathbf{u})\|^2}{2\sigma^2}\right\}. \quad (3)$$

where P is the patch around point considered and $\mathbf{I}(\mathbf{x})$ is the pixel intensity. Here we assumes white, Gaussian noise. For generality, we calculate $\sum_{\mathbf{x}\in P} \|\mathbf{I}(\mathbf{x}) - \mathbf{I}'(\mathbf{x}+\mathbf{u})\|^2$ using the SSD method for each patch of 5×5 pixels, and we enumerate all possible velocities between two frame $\mathbf{I}, \mathbf{I}'$ in the range of $\mathbf{u} \in \{-12, ..., +12\}^2$ pixels.

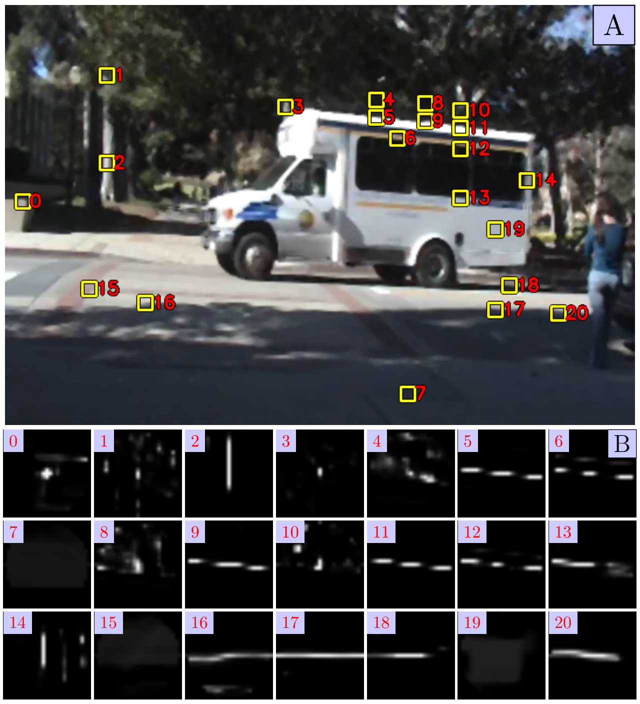

Fig. 3 Posterior probability map of SSD model — (A) patches with numbers; (B) probability map of each numbered patch. Better viewed in color.

Fig. 3(B) shows the full posterior probability maps for 20 typical patches in a video clip in (A). We then compute the local entropy as defined in Definition 2. This is quite time consuming but it is an accurate account of the intrackability. The computation can be accelerated by sub-sampling the velocity vectors or computing SSD in a gradient descent manner. The probability maps in Fig. 3(B) have a large shape variation. There are 4 typical cases

1. Spot shape, for example, 0, 1, and 3; these patches are often corner points and have lowest intrackability;
2. Ridge shape, for example, 2, 14, 16, 17; these patches are edges or ridges and have mid-level intrackability;
3. Multi-modal, for example, 4, 5, 6; these patches are feature points with similar nearby distractors or imperfect edges, and also have mid-level intrackability;
4. Uniform, for example, 7, 15, 19; these patches are often flat regions and have highest intrackability.

In summary, one can see that many of the probability maps cannot be approximated by a simple distribution such as 2D Gaussian.

3 Statistical characteristics of video complexity

This section presents an empirical study on the statistics of natural video clips. We use local intrackability to characterize the video complexity and illustrate the changes of representations over two main axes of changes. Our objective is to gain insight into the various regimes of motion patterns.

3.1 Histograms of local intrackability

As local intrackability is computed in a local space-time volume, we collect a histogram of the intrackabilities by pooling them over the image lattice Λ, following the study of natural image statistics where people collected histograms of local filtered responses.

In our first experiment, we collect a set of 202 video clips of birds from various websites, such as National Geographic and Flickr. Each clip has 6 frames and is resized to the same size (176×144) so that the intrackability is computed in the same range. The reason why we choose bird videos is that birds are captured at a wide range of scales (distance), density, and motion dynamics against clean sky or water. They are ideal for studying the change of representations. As we set $\mathbf{u}$ in the range of $\{-12, ..., +12\}^2$, the maximum value of intrackability is $\log(25 \times 25) \approx 6.4$. We select 60 bins for the histogram of local intrackability and thus treat it as a 60-element vector. To better calculate the distance (i.e., Bhattacharyya distance (Comaniciu et al, 2003)) between histograms, we take the square root of each element.

Fig. 4 shows six typical examples of the square-rooted histograms of local intrackabilities. From our experiments, we observe that there are in general three regimes of motion patterns in these video.

- Flat or noisy videos, such as examples 1 and 6 where the birds are far away and very dense. The intrackability is mostly focused on the right end of the histogram. By flat or noisy videos, we mean the situations where pixel values are almost invariant or dominated by white noise.
- Structured videos, such as example 2, where the birds are close and sparse. The intrackability histogram is widely spread as it contains elements that are trackable (e.g. the corners of bird shapes) and elements that are intrackable (flat patches inside and outside the birds). By structured videos, we mean the situations where discriminable edges and junctions can be spotted.
- Textured video, such as example 4, where the birds are dense but distinguishable from each other. The birds generate texture images of middle granularity. By textured video, we mean the situation where there are many similar appearance features organized as a uniform pattern.

As we zoom-out from example 4 to examples 3, 5, and 6, we gradually observe a clear migration from the low intrackability bins to the high intrackability bins, and finally it will end up like example 1. Row 3 of Fig. 4 verifies our intuitive observation. We conduct a PCA analysis over the 202 square-rooted histograms. The mean histogram has two peaks at two ends. The two eigen-vectors clearly identify the two major transitions. The first eigen-vector shows the change between the trackable (textured or structured, intrackability in $[0, 4.5]$) and the intrackable (flat or noise, intrackability in $[4.5, 6]$), reflecting the increasing complexity. The second eigen-vector shows the change between the highly trackable $[0, 2]$ and the less trackable $[4, 4.5]$, reflecting the change of granularity in scaling. That is, the first axis tell trackable from intrackable and trackable has three cases — highly trackable, middle trackable and less trackable, which are further described by the second axis.

3.2 Scatter plot and variation directions

In our second experiment, we visualize the two types of transitions observed in the previous step. We embed the 202 bird videos in the two dimensions spanned by the two eigen-vectors, and show the result in Fig. 5. We collect the videos on the boundary of the scatter plot and find the two curves representing the two major changes between the most intrackable videos (flat videos on the upper-left corner) and the most trackable (large grained textures on the upper-right corner). We call the flat videos intrackable and large grained textures as trackable, this conflicts with intuition that the flat ones are easier to track. More precisely, the upper-left videos include *objects* that are easier to track, but the videos themselves are not. Before we select which elements to track, we have no idea of objects (suppose we do not have background modeling or object detection). If we try to track all the elements in a video, the intrackabilities of blank areas are higher because of the aperture problem. We need to remove the blank regions, which are both difficult to track and meaningless in most cases. This is the motivation of representation pursuit in Section 4.

Why is this interesting? Traditional vision research on video has been studied in two separate domains: (i) trackable motion including motion flow analysis and object tracking, and (ii) intrackable motion or textured motion. Our experiment shows, perhaps for the first time in the literature, that there is a continuous transition between the two domains. Furthermore, this transition occurs along two axes. The bottom of Fig. 5 visualizes some videos along the two curves. The first row displays videos along the upper boundary of the plot and reflects the change of bird density. The second row displays videos along the lower boundary and reflects the change of bird granularity through scaling. The videos in the interior of the plot in Fig. 5 contain birds of different

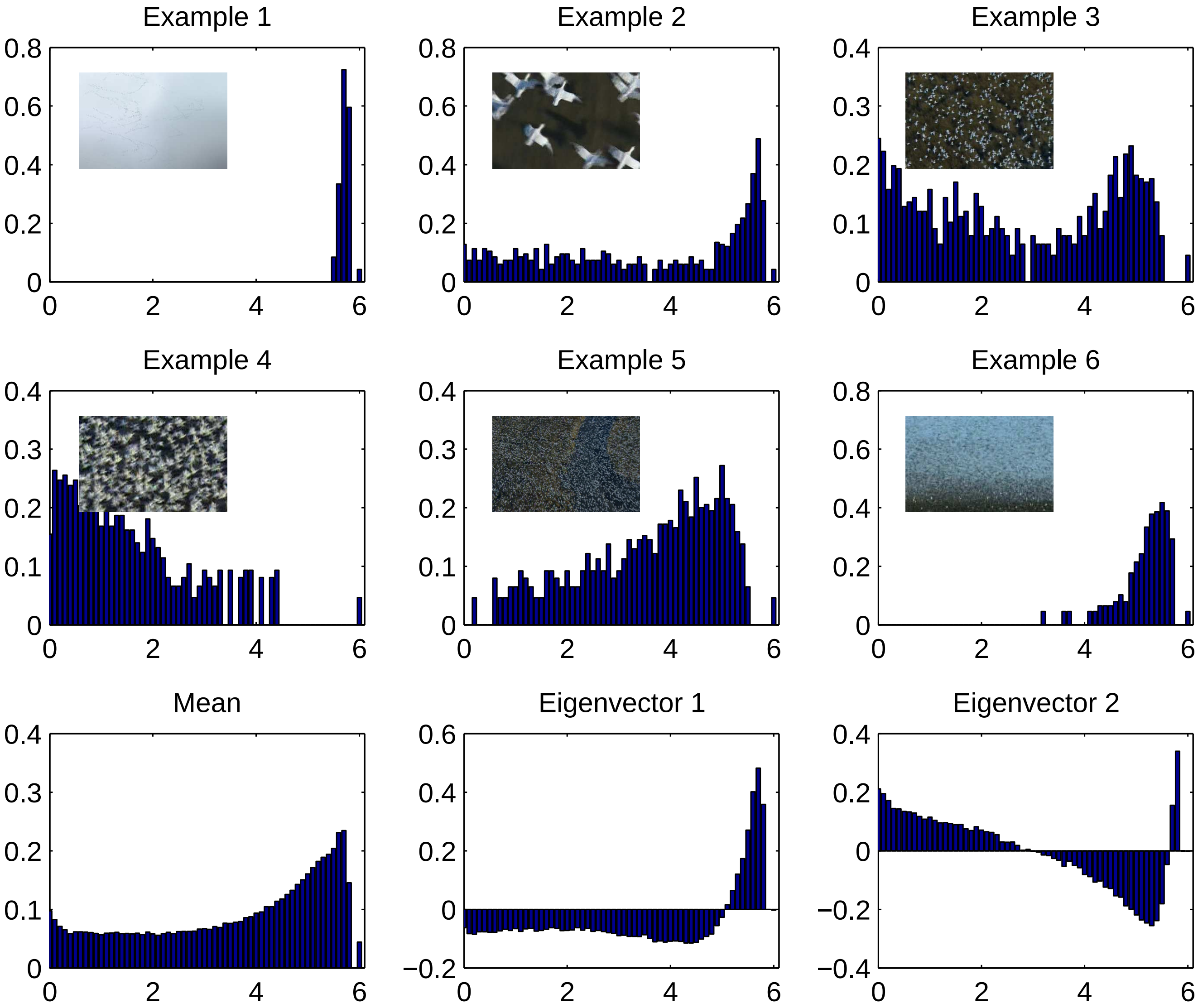

Fig. 4 (Row 1-2) Six examples of the square-rooted histograms of local intrackabilities. (Row 3) Three components are fit to the mean histogram, and the first two eigen-vectors of these square-rooted histograms reveal the transitions between the three components.

sizes and numbers and therefore are mixtures of the ones on the boundary. Such observations call for a unified framework for modeling all video patterns and for a continuous transition between the various motion representations.

As the two curves form a loop of two continuous changes, we re-organize the videos on the boundary and visualize them in Fig. 8.

For tracking tasks, we are interested in trackable elements in a video and most intrackable areas are discarded to reduce computing burden. We apply a threshold ($1/3$ of the maximal intrackability value) on each video to obtain a set of trackable elements, and the sum of the intrackabilities of all trackable elements in a video provides the uncertainty of the tracking task. Fig. 9 plots the total sum of the intrackabilities in these trackable areas for all the videos on the blue curve and red curve in Fig. 5. This figure illustrates that the sum achieves the peak at populous videos, which means that they are the most difficult to track when we have discarded the intrackable uniform regions and textured regions with high intrackabilities. For videos with modest number of objects, each feature point has less ambiguity. For flat or noisy videos, the number of trackable points is almost zero, so the tracking algorithm can do nothing. Therefore, it has to switch to appearance models, such as the spatio-temporal auto-regression (STAR) model, to represent the video appearance without explicitly computing the motion. In this sense, intrackability is indeed a good measure for the transition of models.

In our third experiment, we extend the study of bird video to general natural video clips in the same way. We collected a set of 237 video clips containing a large variety of objects, such as people, birds, animals, grass, trees, water with

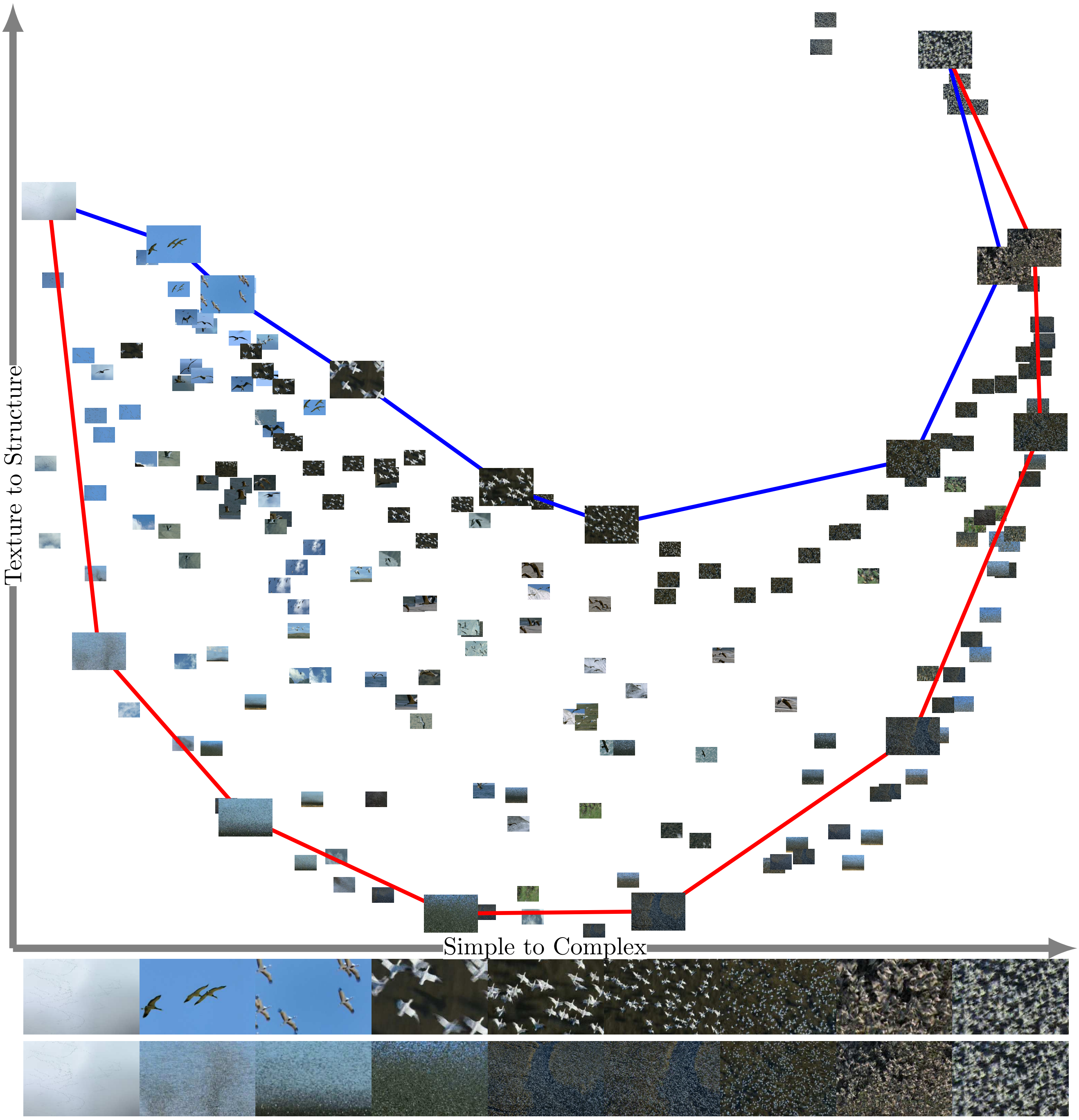

Fig. 5 PCA embedding of histograms of intrackabilities for the 202 bird videos in two dimensions. Red and blue curves show two typical transitions: The blue curve (top) shows density changes of elements (objects) in the video: from a few birds to thousands of birds. The red curve (bottom) shows scales changes in the videos: from fine granularity to large granularity. In the bottom, the first row shows the video examples on the blue curve and the second row shows the video examples on the red curve.

different speed and density in natural environments. Fig. 10 shows the results of the two-dimensional embedding.

The result coincides with the bird experiments. The 237 video clips are bounded by the two typical transition curves. The bottom of Fig. 10 shows the typical video clips along the two curves.

For comparison, we use Shi-Tomasi texturedness measure and Harris-Stephen R score to do the same PCA. The results are shown in Fig. 6 and 7 respectively. Shi-Tomasi is a local measure that only accounts for the gradient information in the patch, and does not take into account similar objects in the surrounding neighborhood. Therefore, it takes videos of dense small objects as trackable, and puts them on

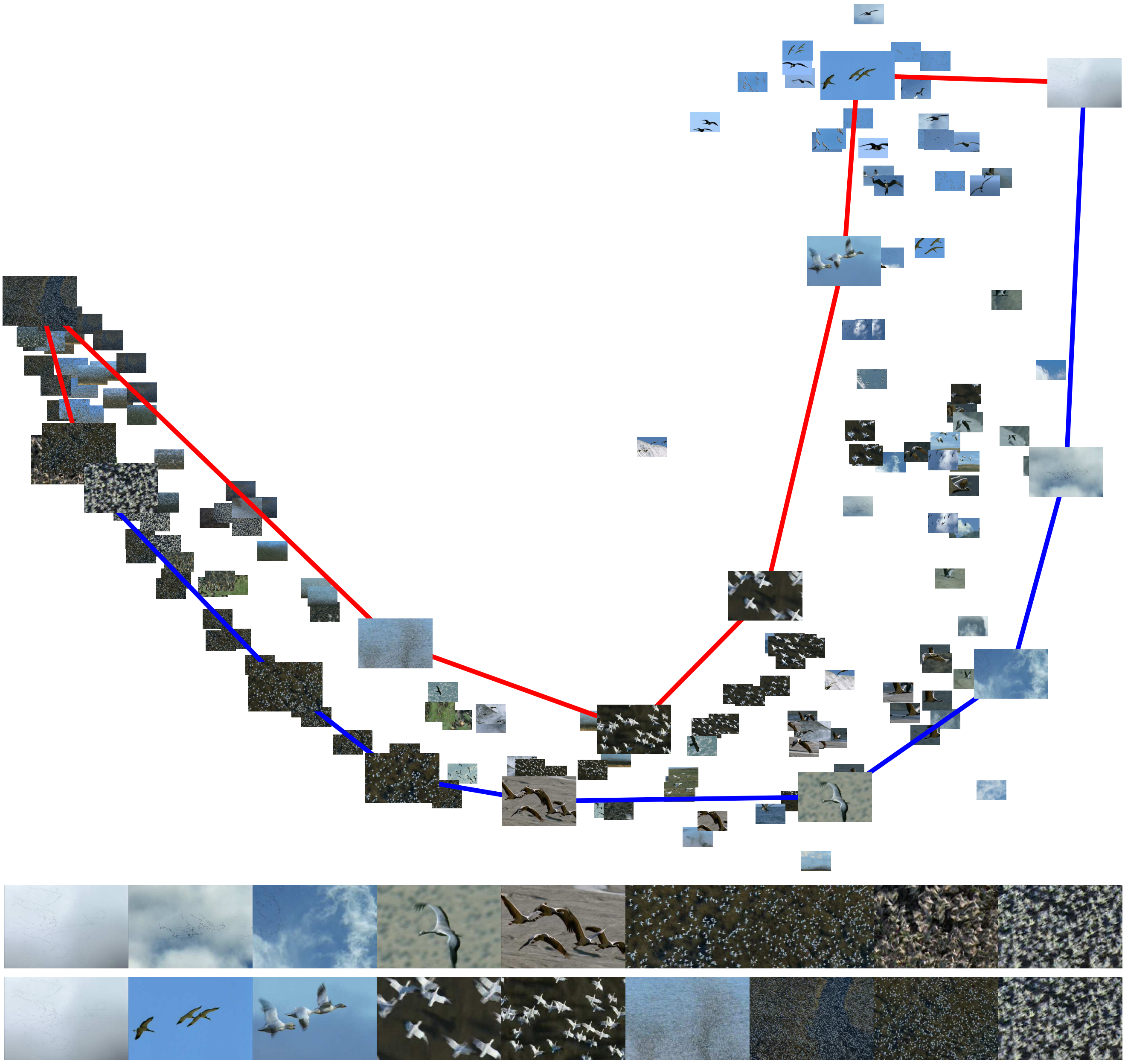

Fig. 6 PCA embedding of histograms of Shi-Tomasi texturedness measure for the 202 bird videos in two dimensions. Red and blue curves enclosing the region are not as reasonable as in Fig. 5.

the left side of Fig. 6. In the results of the Harris-Stephens R score (Fig. 7), the structural videos are concentrated in a small region near the right side.

4 Pursuing hybrid video representations

In this section, we study a method for automatically selecting the optimal video representations based on an intrackability measure. We start with an overview of some popular representations in four different regimes.

4.1 Overview of four video representations

We have discussed the four distinct representations in Fig. 1: contour, kernel or PCA Basis, dense motion field, and joint image appearance model. We divide them into two categories. For the first three types of representations, there are a number of elements to track, so we call them the trackable motion. We denote the appearance and geometry of these elements by a dictionary $\Delta = \{\psi_1, ..., \psi_n\}$ and their motion velocity by $W = (\mathbf{u}_1, ..., \mathbf{u}_n)$. For the fourth representation, there is nothing to track and thus W does not contain

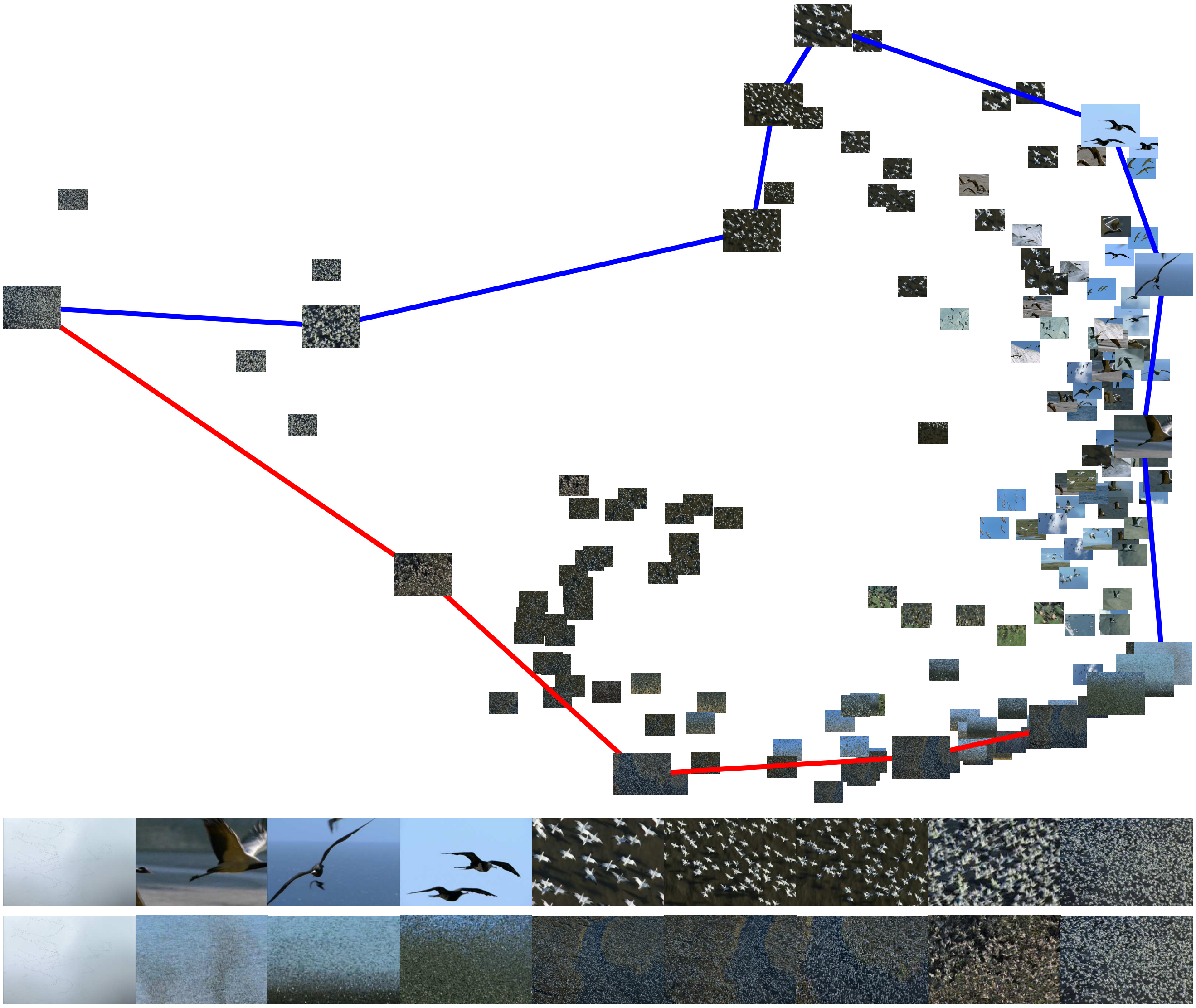

Fig. 7 PCA embedding of histograms of Harris-Stephens R score for the 202 bird videos in two dimensions. Red and blue curves show two typical transitions on the boundary. The lower curve is the same as in Fig. 5. But the upper curve is not as reasonable as in Fig. 5. Additionally, the hollow near the upper curve makes use difficult to determine its real boundary.

velocity variables and only has some parameters. We call it intrackable motion.

Trackable motion For the contour, kernel, PCA basis, and dense motion, the posterior probability is

$$p(W|\mathbf{I}, \mathbf{I}'; \Delta) \propto p(\mathbf{u}_1, \cdots, \mathbf{u}_n) \prod_{i=1}^{n} p(\mathbf{I}_{\Lambda_i}|\mathbf{u}_i, \mathbf{I}'; \psi_i) \qquad (4)$$

In the above formula, $p(\mathbf{I}_{\Lambda_i}|\mathbf{u}_i, \mathbf{I}'; \psi_i)$ is the local likelihood probability for tracking an element ψ_i in a patch (domain) Λ_i discussed before,

$$p(\mathbf{I}_{\Lambda_i}|\mathbf{u}_i, \mathbf{I}'; \psi_i) \propto \exp\left\{-\frac{\sum_{\mathbf{x}\in P_i} \|\mathbf{I}(\mathbf{x}) - \mathbf{I}'(\mathbf{x}+\mathbf{u}_i)\|^2}{2\sigma^2}\right\}$$

which is consistent with Eq. 3 if we assume a uniform motion prior. For clarity and generality, we use the SSD measure based on the image patch $\mathbf{I}(\mathbf{x})$ and $\mathbf{I}'(\mathbf{x}+\mathbf{u}_i)$ for $\mathbf{x} \in P_i$, this could be replaced by other features defined on $\psi_i(\mathbf{x})$ and $\psi'_i(\mathbf{x}+\mathbf{u})$.

The joint probability $p(\mathbf{u}_1, \cdots, \mathbf{u}_n)$ is a contextual model for the coupling of these moving elements.

- In contour tracking (Maccormick and Blake, 2000; Sato and Aggarwal, 2004; Black and Fleet, 2000), all the points may show a rigid affine transform plus some local small deformations. Furthermore the velocity $\mathbf{u}_i = (u_i^{\perp}, u_i^{\|})$ is reduced to $u_i^{\perp}$ containing only the direction perpendicular to the contour. The tangent speed is discarded as it cannot be inferred reliably (due to high entropy). The element ψ_i could be the patch or image profile along the normal direction of the contour at key points.

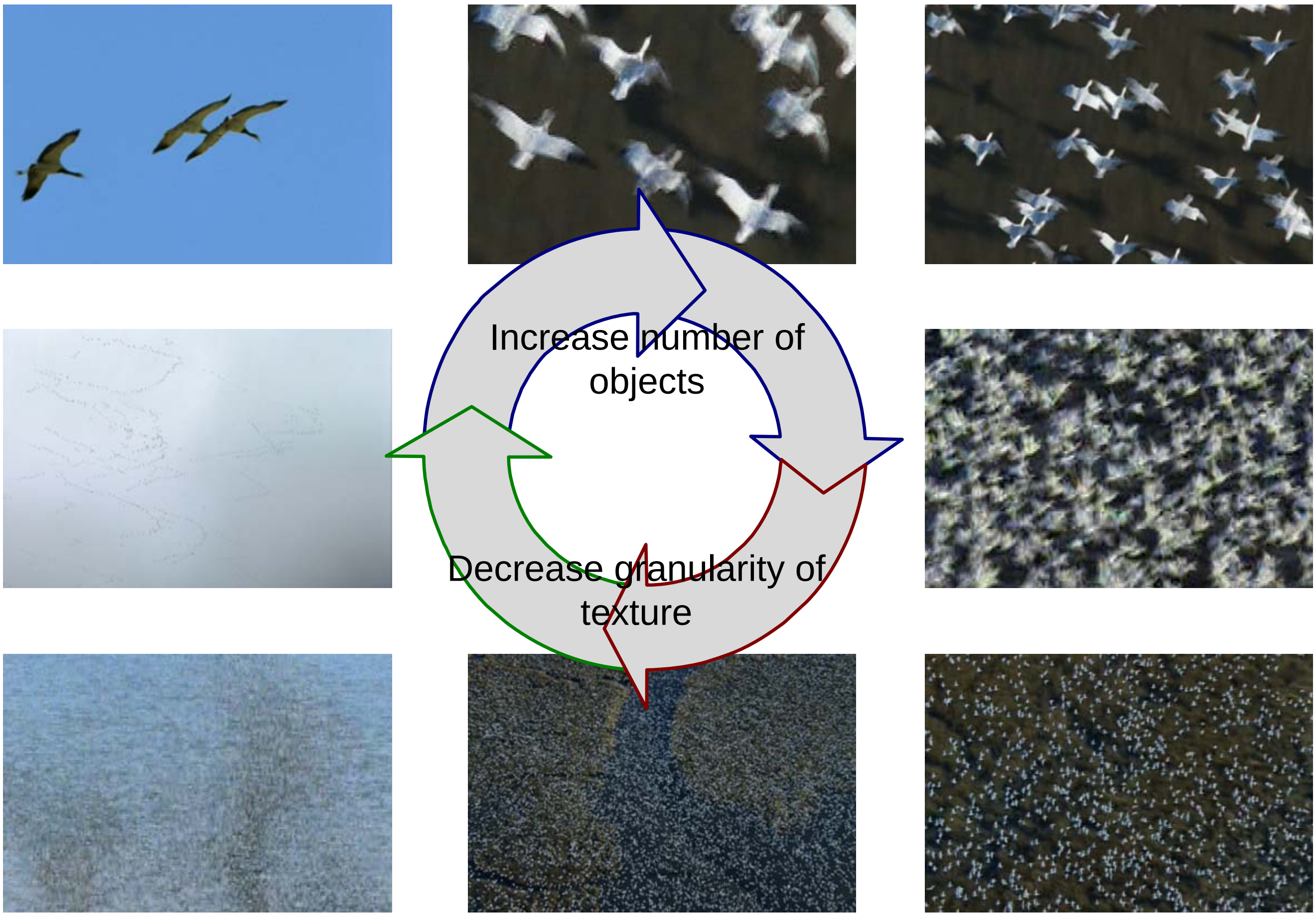

Fig. 8 The continuous change between different videos through two major axes:the change of density and the change of granularity.

- In kernel tracking (Comaniciu et al, 2003; Collins, 2003), a kernel in the shape of an ellipse, rectangle or other geometric primitive is defined for an object, and all the interior feature points are assumed to have the same velocity (rigid) or adjacent points are assumed to have similar velocity. The element ψ_i could be a feature descriptor like SIFT or PCA basis.
- In the dense motion field, $(\mathbf{u}_1, \cdots, \mathbf{u}_n)$ is regulated by a Markov random field (Horn and Schunck, 1981; Black and Fleet, 2000). The element ψ_i is either a pixel or a feature point.

These models $p(\mathbf{u}_1, \cdots, \mathbf{u}_n)$ essentially reduce the randomness of the motion or equivalently the degrees of freedom in W. In the next subsection, we will pursue such representations by reducing the variables in W.

Intrackable motion When the motion includes a large number of indistinguishable elements, it is called dynamic texture (Szummer and Picard, 1996; Fitzgibbon, 2001; Soatto et al, 2001) or textured motion (Wang and Zhu, 2003), such as fire flame, water flow, evaporating steam etc. As the moving elements are indistinguishable, the velocity cannot be inferred meaningfully and W is empty. These videos are represented by appearance models directly, typically by regression models. An example is the spatio-temporal autoregression (STAR) model,

$$\mathbf{I}(\mathbf{x}, t) = \sum_{(\mathbf{y},s)\in\partial(\mathbf{x},t)} \alpha_{\mathbf{y}-\mathbf{x},s-t}\mathbf{I}(\mathbf{y}, s) + n(\mathbf{x}, t), \ \forall \mathbf{x}, t. \quad (5)$$

That is, the pixel intensity at $\mathbf{x}$ and frame t is a regression of other pixels in the spatio-temporal neighborhood ($\partial(\mathbf{x}, t)$) plus some residual noise $n(\mathbf{x}, t)$. The model is represented by parameters $\Theta = (\alpha_{\mathbf{y}-\mathbf{x},s-t})$ which are often homogeneous in space and time, learned by fitting certain statistics. The size of the spatio-temporal neighborhood may be selected for different videos. In general, one can rewrite the video $\mathbf{I}_\Lambda[0, T]$ in a Gaussian Markov random field model,

$$p(\mathbf{I}_\Lambda[0, T]; \Theta) \propto \exp\left\{-\frac{\sum_{t=1}^{T}\sum_{\mathbf{x}\in\Lambda} n^2(\mathbf{x}, t)}{2\sigma_o^2}\right\}. \quad (6)$$

4.2 Automatic selection of hybrid representations

A natural video often includes multiple objects or regions of different scales and complexities and thus is best represented

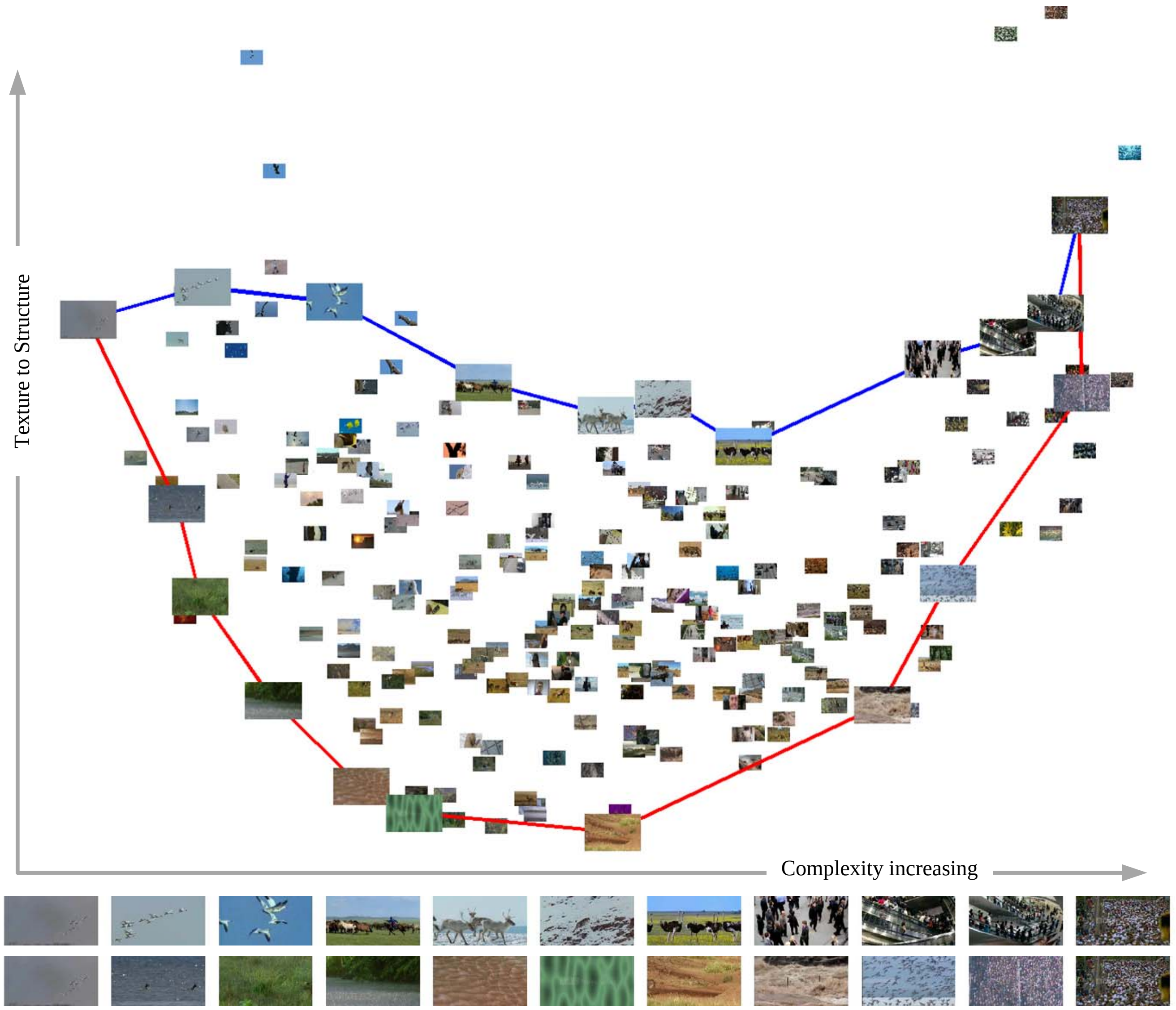

Fig. 10 PCA embedding of histograms of intrackabilities for the 237 natural videos in two dimensions. Red and blue curves show two typical transitions: The blue curve (top) shows density changes of elements (objects) in the video. The red curve (bottom) shows scales changes in the videos: from fine granularity to large granularity. In the bottom, the first row shows the video examples on the blue curve and the second row shows the video examples on the red curve.

by a hybrid representation. Fig. 11 shows an example. The bird in the foreground is imaged at a near distance. Some spots (the head, the neck, the leg, and the tips of the wings) are distinguishable from the surrounding areas and therefore their intrackability is low as shown in (b). They should be represented by key points or kernels that can be tracked over a number of frames. The points along the bird outline are less trackable and have higher intrackability value in (b). But after projecting to line segments through merging adjacent points and dropping the tangent directions from W, these line segments become trackable. Fig. 11(c) shows the intrackability map of the lines. For the remaining areas, the wavy water in the background is textured motion and the interior of the bird is flat area. These are intrackable, and thus are represented by STAR (or MRF) models. The so-called tri-map in (d) illustrates the three different regimes of models calculated according to their intrackabilities. This representation will have to change as the bird flies close to or away from the camera, or as the number of birds changes, as many other videos have shown in the previous section.

Automated selection and on-line adaptation of such hybrid representations is of practical value for both computer and biological visual systems. Given the limited resources (memory and computing capacity), the system must perform a trade-off between more detail and less intrackability wisely. Psychological experiments show that human vision changes the task and perception as well when the complexity exceeds the system capacity (Pylyshyn, 2004, 2006).

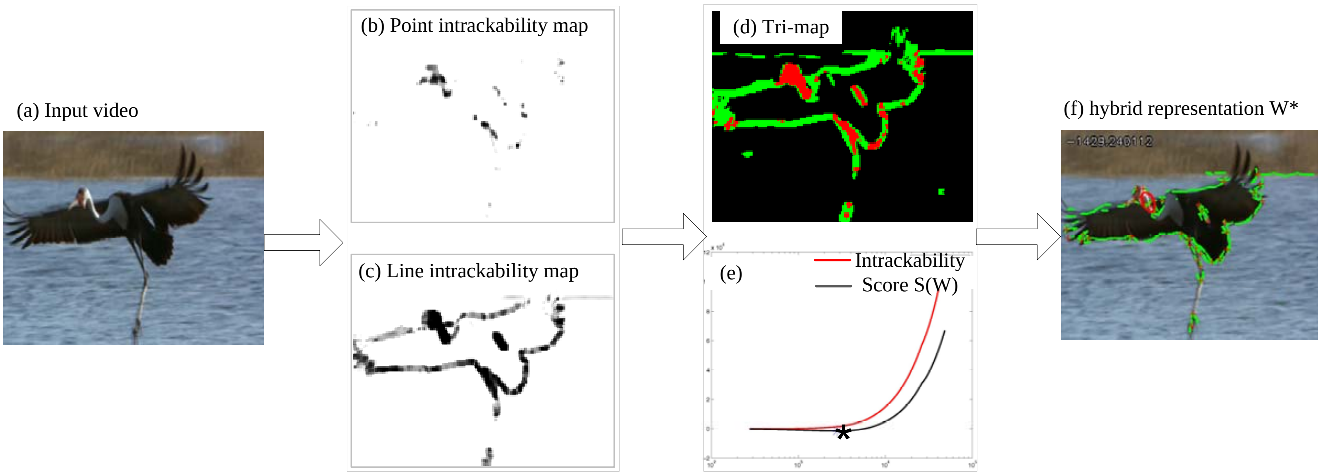
(a) Input video
(b) Point intrackability map
(c) Line intrackability map
(d) Tri-map
(e)
Intrackability
Score S(W)
(f) hybrid representation W*
-1429.240112

Fig. 11 Pursuing a hybrid video representation. From an input video (a), we compute the intrackability map (b) and projected line intrackability map (c) where darker points have lower intrackability. The trimap (d) visualizes the three different representations: red spots are trackable and represented by key points or kernels; green areas are trackable after projecting to line segments and therefore are represented by contours, and the black area is intrackable motion and is represented by STAR models. We plot the intrackability $\mathcal{H}(W)|\mathbf{I},\mathbf{I}'$ and $S(W)$ in (e) where the horizontal axis is the number of variables in W from simple to complex. The optimal representation W^* (f) corresponds to the minimum cost $S(W)$ shown by the star point on the curve in (e).

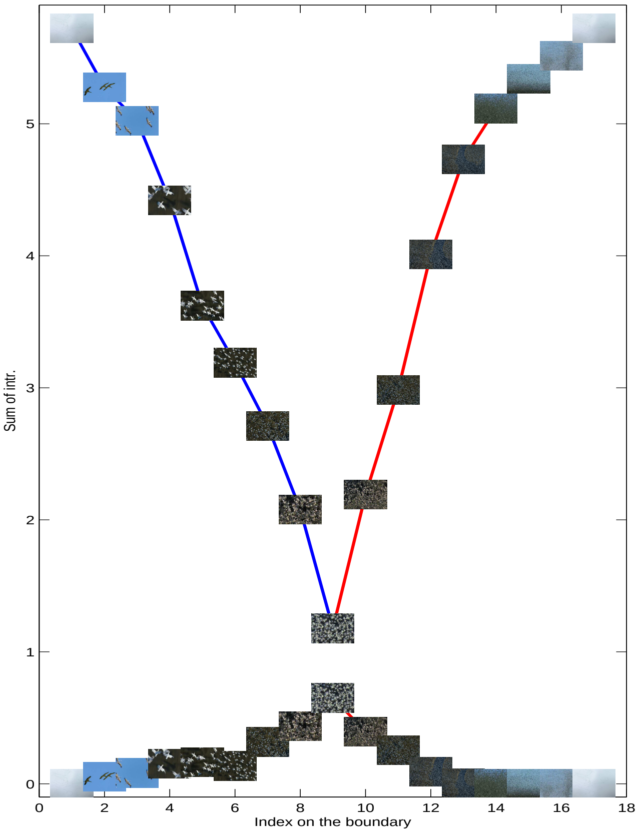

Fig. 9 Total intrackabilities in entire video and trackable area for each video on the boundary. Top curve is the sums of intrackabilities in entire videos; bottom curve is the sums of intrackabilities in trackable areas. The red and blue curves correspond to those in Fig. 5.

The criterion that we use for selecting the hybrid representation W^* includes two objectives:

– The representation should be as detailed as possible so that it does not miss important motion information. This encourages representation with high complexity.
– The representation should be inferred reliably. In other words, it has a lower uncertainty or entropy.

The two objectives are combined into the following function,

$$S(W) = \mathcal{H}\{W|\mathbf{I}_\Lambda[t, t+\tau]\} - A(W). \tag{7}$$

We assume W is fixed in a short duration τ, $\mathcal{H}\{W|\mathbf{I}_\Lambda[t, t+\tau]\}$ is the instance intrackability defined before, and $A(W)$ is the description (coding) length for the variables in W. We minimize the criterion $S(W)$ to obtain the best representation, $W^* = \arg\min_W S(W)$.

Fig. 11(e) gives an example of the criterion $S(W)$ against the number of variables in W. By minimizing this function, we obtain a representation W^* which is shown in Fig. 11(f). It consists of a number of trackable points, lines, contours and intrackable regions.

MAP is a popular method for video representation, e.g., (Wang et al, 2005; Wang and Zhu, 2008). Video representation can be decomposed into two sub problems, 1) choosing variables and 2) estimating the values of the selected variables. The MAP work in fact addresses both of these with a single criterion. In this paper, we encourage separate investigation of the two and focus on the first problem, which is more important. Our answer to the first problem is to select what are good for the second problem. After the variables are determined, the estimation of their values can be accomplished by MAP, expectation or sampling.

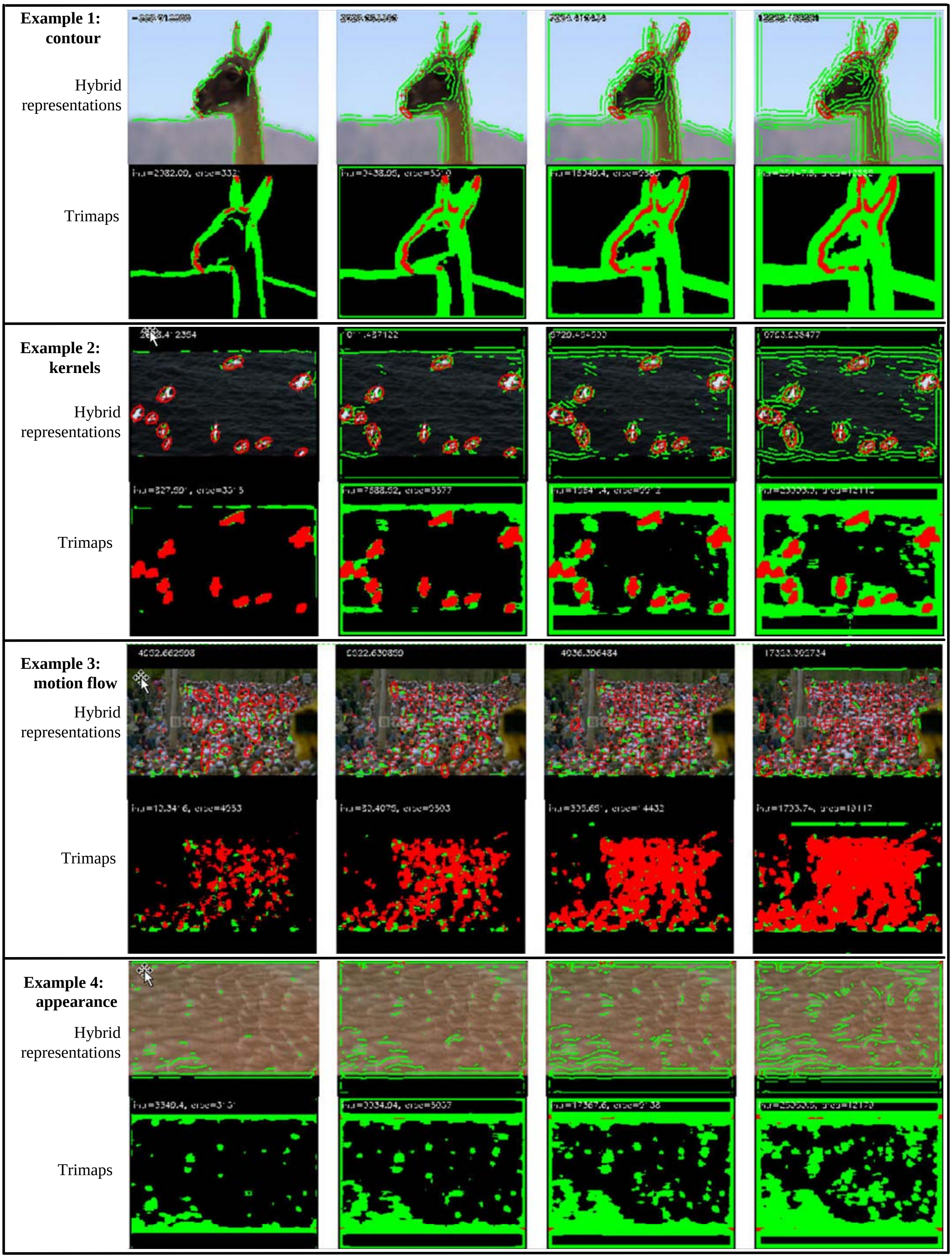

Fig. 12 Trimaps and pursued hybrid representations at different thresholds: red — trackable points, green — trackable lines in projected direction, black — intrackable points. For each video, from left to right, the threshold varies from high to low. The first video can be best represented by contours. The second video can be best represented by kernels. The third video can be best represented by dense points. The fourth can be best represented by appearance models.

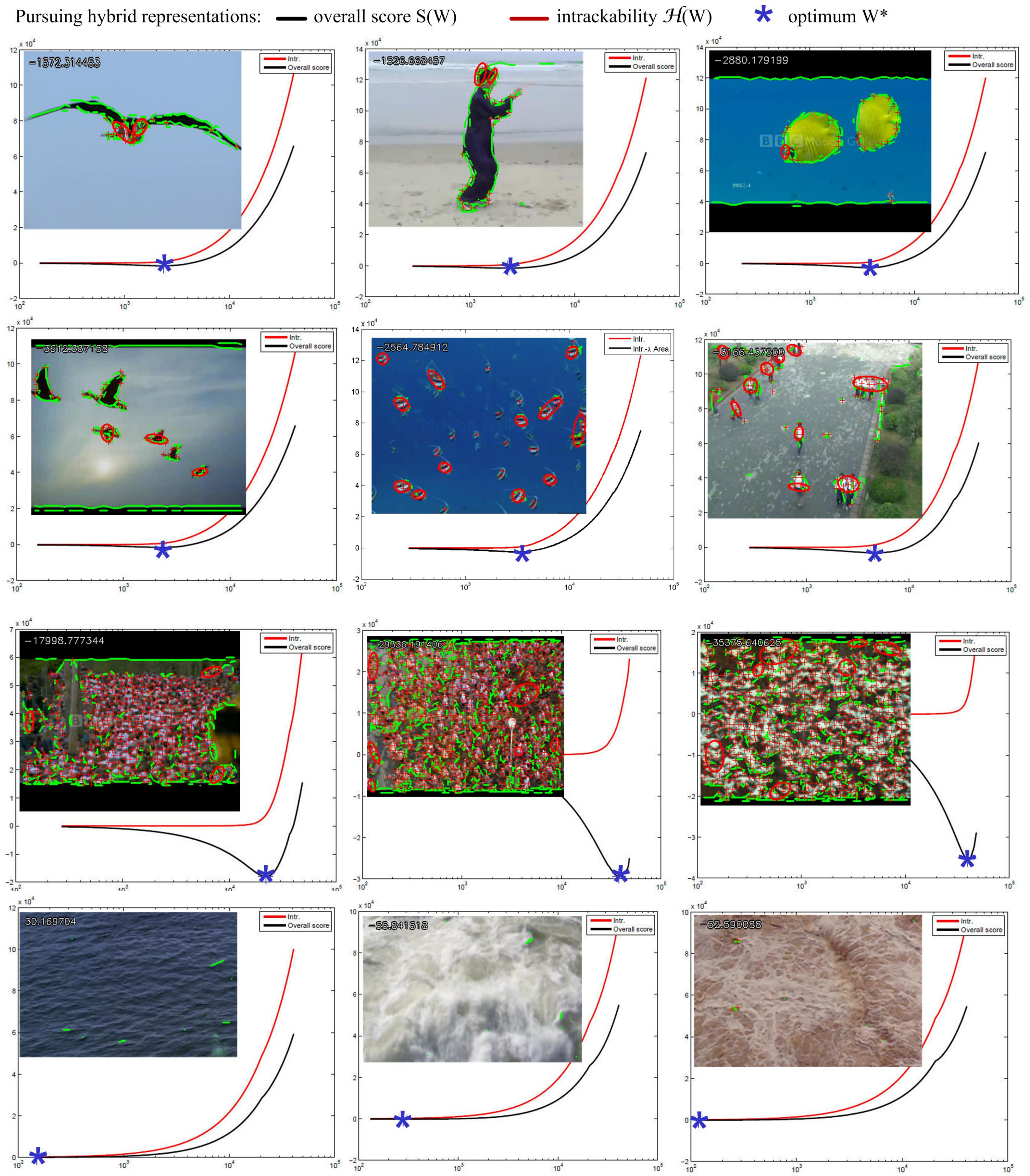

Fig. 13 More results of hybrid representation pursuit in 12 video clips. In each example, we show the hybrid representations: red crosses are trackable points, red ellipses are grouped kernels; and green curves are the trackable contours. In the background, we show the cost curves $S(W)$ in black and the intrackability curve in red. The asterisks on the black curves indicate the minima. The horizontal axis is the number of variables in W. The vertical axis is the intrackability or the cost.

In the following, we introduce the representation projection operators that compute W^* and realize the transition between the models.

4.3 Representation projection

We start with an overly detailed representation $W_o = (\mathbf{u}_1, ..., \mathbf{u}_N)$ with N being the number of points that are densely sampled in the image lattice. The motion velocities $\mathbf{u}_i, i = 1, 2, ..., N$ are assumed to be independent in the range of $[-12, 12]^2$ pixels. Therefore we have

$$S(W_o) = \sum_{i=1}^{N} \mathcal{H}\{\mathbf{u}_i|\mathbf{I}, \mathbf{I}'\} - \lambda \cdot 2N,$$

where λ is the description length of each velocity direction. W_o is the most complex representation, corresponding to the right end of the plot in Fig. 11(e). We convert it to a hybrid representation W^* by representation projection with four types of operators. Each operator will reduce $S(W_o)$ in a greedy way (i.e. pursuit).

1. Point dropping. We may drop the highly intrackable points (or image patches). By dropping an element $\mathbf{u}_i$ from W, the change of $S(W)$ is

$$\Delta_i = -\mathcal{H}\{\mathbf{u}_i|\mathbf{I}, \mathbf{I}'\} + 2\lambda \ < \ 0.$$

In other words, points with $\mathcal{H}\{\mathbf{u}_i|\mathbf{I}, \mathbf{I}'\} < 2\lambda$ remain in W as “trackable points” which are indicated by the red crosses in Fig. 11(f). We also perform a local-non-maximum suppression. Because our local intrackability is estimated based on patches, any points within a neighborhood (say 5×5) of the trackable points will be suppressed.

2. Velocity projection. For the remaining points, we project the velocity $\mathbf{u}$ to one dimension $u_\perp$ so that the projected velocity has the lowest intrackability,

$$\mathcal{H}\{u_\perp|\mathbf{I}, \mathbf{I}'\} = \min_{\xi} \mathcal{H}\{\langle \xi, \mathbf{u}\rangle|\mathbf{I}, \mathbf{I}'\}$$

in which ξ is a unit vector representing the selected orientation. If the patch contains an edge, the most likely orientation ξ is the normal direction of the edge. Fig. 11(c) illustrates the projected intrackability. If we let u' be the component of $\mathbf{u}$ that is perpendicular to $u_\perp$, that is $\mathbf{u} = (u_\perp, u')$. Then we have

$$\begin{aligned} \mathcal{H}\left\{\mathbf{u}|\mathbf{I}, \mathbf{I}'\right\} &= \mathcal{H}\left\{(u_\perp, u')|\mathbf{I}, \mathbf{I}'\right\} && (8) \\ &= \mathcal{H}\left\{u_\perp|\mathbf{I}, \mathbf{I}'\right\} + \mathcal{H}\left\{u'|u_\perp, \mathbf{I}, \mathbf{I}'\right\} && (9) \end{aligned}$$

in which $\mathcal{H}\left\{u'|u_\perp, \mathbf{I}, \mathbf{I}'\right\}$ is the conditional entropy of u' given $u_\perp$, and is always non-negative. Therefore we have

Proposition 1 *Intrackability decreases with representation projection, i.e.,* $\mathcal{H}\{u_\perp|\mathbf{I}, \mathbf{I}'\} \leqslant \mathcal{H}\left\{\mathbf{u}|\mathbf{I}, \mathbf{I}'\right\}$.

While $\mathbf{u}$ is intrackable, its component $u_\perp$ may still be trackable along the normal direction. Thus, we replace the element $\mathbf{u}_i$ by $u_\perp$ in W. This leads to a change of $S(W)$:

$$\Delta_i = \mathcal{H}\{\mathbf{u}_i|\mathbf{I}, \mathbf{I}'\} - \mathcal{H}\{u_\perp|\mathbf{I}, \mathbf{I}'\} + \lambda \ < \ 0.$$

In other words, we drop the direction which has large entropy.

By thresholding the intrackability map and projected intrackability map, we obtain a trimap showing the trackable, trackable in one direction and intrackable regions. Fig. 11(d) shows a dense trimap where a red point is trackable, a green point is trackable in a projected direction, and a black point is intrackable. Fig. 12 shows the trimaps for four examples with different choices of thresholds.

3. Pair linkage. After eliminating the points in the previous two steps, we further reduce $S(W)$ by exploring the dependency between the elements. We sequentially link adjacent points or lines into a chain structure (contours). Suppose the resulting contour has k points/lines $(\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_k)$, we assume these elements follow a Markov chain, so

$$p(\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_k|\mathbf{I}, \mathbf{I}') = p(\mathbf{u}_1|\mathbf{I}, \mathbf{I}') \prod_{i=2}^{k} p(\mathbf{u}_i|\mathbf{u}_{i-1}, \mathbf{I}, \mathbf{I}').$$

Proposition 2 *Pair linking reduces the intrackability*

$$\begin{aligned} \mathcal{H}\{\mathbf{u}_1, ..., \mathbf{u}_k|\mathbf{I}, \mathbf{I}\} &= \sum_{i=1}^{k} \mathcal{H}\{\mathbf{u}_i|\mathbf{I}, \mathbf{I}'\} - \sum_{i=2}^{k} \mathcal{M}(\mathbf{u}_i, \mathbf{u}_{i-1}|\mathbf{I}, \mathbf{I}') \\ &\leqslant \sum_{i=1}^{k} \mathcal{H}\{\mathbf{u}_i|\mathbf{I}, \mathbf{I}'\}, \qquad (10) \end{aligned}$$

where $\mathcal{M}(\mathbf{u}_i, \mathbf{u}_{i-1}|\mathbf{I}, \mathbf{I}') \geqslant 0$ *is the conditional mutual information between two adjacent elements.*

The mutual information is defined as

$$\begin{aligned} &\mathcal{M}(\mathbf{u}_i, \mathbf{u}_{i-1}|\mathbf{I}, \mathbf{I}') && (11) \\ &= \sum_{\mathbf{u}_i, \mathbf{u}_{i-1}} p(\mathbf{u}_i, \mathbf{u}_{i-1}|\mathbf{I}, \mathbf{I}') \log \frac{p(\mathbf{u}_i, \mathbf{u}_{i-1}|\mathbf{I}, \mathbf{I}')}{p(\mathbf{u}_i|\mathbf{I}, \mathbf{I}')p(\mathbf{u}_{i-1}|\mathbf{I}, \mathbf{I}')} && (12) \\ &= \mathcal{H}\{\mathbf{u}_i|\mathbf{I}, \mathbf{I}'\} - \mathcal{H}\{\mathbf{u}_i|\mathbf{u}_{i-1}, \mathbf{I}, \mathbf{I}'\} && (13) \end{aligned}$$

Eq. (12) shows that it is Kullback-Leibler divergence from $p(\mathbf{u}_i, \mathbf{u}_{i-1}|\mathbf{I}, \mathbf{I}')$ to $p(\mathbf{u}_i|\mathbf{I}, \mathbf{I}')p(\mathbf{u}_{i-1}|\mathbf{I}, \mathbf{I}')$, and therefore non-negative.

In $S(W)$, the reduction of the intrackability is the mutual information at each step, the number of variables $A(W)$ remains the same, though we may need to index the chain structure with a coding length of ϵ. So each time by linking a pair of elements $\mathbf{u}_i$, we have a change of $S(W)$ by

$$\Delta_i = -\mathcal{M}(\mathbf{u}_i, \mathbf{u}_{i-1}|\mathbf{I}, \mathbf{I}') + \epsilon < 0. \qquad (14)$$

We compute $\mathcal{M}(\mathbf{u}_i, \mathbf{u}_{i-1}|\mathbf{I}, \mathbf{I}')$ by Eq. (13). To compute the conditional entropy $\mathcal{H}\{\mathbf{u}_i|\mathbf{u}_{i-1}, \mathbf{I}, \mathbf{I}'\}$, one may enumerate all possible combinations of $(\mathbf{u}_i, \mathbf{u}_{i-1})$, then compute the conditional probability, joint probability and entropy. As a faster approximation, we find the optimal solution $\mathbf{u}^*_{i-1}$ first, and then compute $\mathcal{H}\{\mathbf{u}_i|\mathbf{u}^*_{i-1}, \mathbf{I}, \mathbf{I}'\}$. T-junctions can be found automatically when we greedily grow the set of projected trackable elements by pair linking.

4. Collective grouping. This operator is to group a number of adjacent elements in an ellipse simultaneously into a kernel representing a moving object. Given the velocity $\mathbf{u}_0$ of the kernel, the grouped elements $\mathbf{u}_1, ..., \mathbf{u}_k$ are assumed to be conditionally independent,

$$p(\mathbf{u}_0, \mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_k|\mathbf{I}, \mathbf{I}') = p(\mathbf{u}_0|\mathbf{I}, \mathbf{I}') \prod_{i=2}^{k} p(\mathbf{u}_i|\mathbf{u}_0, \mathbf{I}, \mathbf{I}').$$

Therefore the change of $S(W)$ is

$$\Delta_{1..k} = \mathcal{H}\{\mathbf{u}_0|\mathbf{I}, \mathbf{I}'\} - \sum_{i=1}^{k} \mathcal{M}(\mathbf{u}_i, \mathbf{u}_0|\mathbf{I}, \mathbf{I}') < 0$$

In practice, we place an ellipse around each trackable point in the trimap, and if it contains a few trackable points, for which the best estimations of velocities are very close, then we group them into a kernel.

4.4 Experiment on pursuing hybrid representation

The precise optimization of $S(W)$ is computationally intensive, so we use a greedy algorithm which starts with the dense point representation W_o, then sequentially applies the four operators to reduce $S(W)$. The final result is a hybrid representation consisting of: trackable points (red crosses), trackable lines (green), contours (green), kernels (red ellipses), and the remaining intrackable regions.

In addition to the results in Fig. 11 and 12, we tested the pursuit algorithm on a variety of video clips. Fig. 13 shows 12 examples representing videos of different complexities. In row 1: the foreground objects (bird, human, and fish) exhibit high resolution in a flat background. The contours and short lines dominate the representation. In row 2: the objects (birds, fish, and people) exhibit low resolution and are well separated from the background. Thus, they are represented by kernels. In row 3, the objects (still people, fish, birds) exhibit low resolution and high density. As many elements are still distinguishable in their neighborhood, they are represented by dense trackable points. In row 4, there are no trackable elements; the video becomes a texture appearance and thus described by STAR model.

From the final pursuit results, one can see that most of the feature points and the object contours are captured successfully. The junctions on car (especially the window corner) and person (cloth corners) are well classified as sparse feature points, and the edges and contours are well classified as lines. The horizontal line between the water and sand in the first row is not selected as a trackable line due to weak edge contrasts and similar lines in the neighborhood.

Fig. 14 shows additional results on two longer sequences. The top row shows a swimming shark represented by contour and feature points. The bottom row shows a moving camera approaching a car. At first, the car is very far away, and appears as a feature point. As the camera approaches, it is represented by a kernel. As the camera approaches further, more details are revealed, and it is represented by a set of contours, kernels and feature points.

Fig. 15 shows comparisons with trimaps of the Harris-Stephens corner detector. One can see that, when no similar distraction is present nearby, the results of Harris-Stephens are very similar to ours. But when objects present at smaller scale, Harris-Stephens reports more possible edges. For textured video like water, Harris-Stephens fails to report them as intrackable, while intrackablity succeed because it takes into account the similar distractions nearby. The Harris-Stephens trimaps are determined by tuning parameters such that results for large scale objects are similar to ours.

5 Comparison with other tracking criteria

In this section, we compare the intrackability with two other measures for robust tracking, namely the Shi-Tomasi texturedness measure and the conditional number.

5.1 Intrackability and the Shi-Tomasi texturedness measure

(Shi and Tomasi, 1994) proposed a texturedness criterion for good points to track in two frames. To compare with this criterion, we rewrite the local posterior probability for a point velocity $\mathbf{u} = (u_x, u_y)$ that we discussed before,

$$p(\mathbf{u}|\mathbf{I}, \mathbf{I}') \propto \exp\left\{-\frac{\sum_{\mathbf{x}\in P} |\mathbf{I}(\mathbf{x}) - \mathbf{I}'(\mathbf{x}+\mathbf{u})|^2}{2\sigma^2}\right\}.$$

As is common in optical flow computation, one assumes the image is differentiable with $(\mathbf{I}_x, \mathbf{I}_y)$ being the image gradient. By Taylor expansion we have

$$\mathbf{I}'(\mathbf{x}+\mathbf{u}) = \mathbf{I}(\mathbf{x}) + u_x\mathbf{I}_x + u_y\mathbf{I}_y. \tag{15}$$

Then we can rewrite $p(\mathbf{u}|\mathbf{I}, \mathbf{I}')$ in a Gaussian form,

$$p(\mathbf{u}|\mathbf{I}, \mathbf{I}') = \frac{1}{2\pi \mathrm{det}^{1/2}(\Sigma)} \exp\{-\frac{1}{2}\mathbf{u}\Sigma^{-1}\mathbf{u}'\}. \tag{16}$$

where the inverse covariance matrix is,

$$\Sigma^{-1} = \begin{pmatrix} \sum_{\mathbf{x}\in P} \mathbf{I}_x^2(\mathbf{x}) & \sum_{\mathbf{x}\in P} \mathbf{I}_x(\mathbf{x})\mathbf{I}_y(\mathbf{x}) \\ \sum_{\mathbf{x}\in P} \mathbf{I}_x(\mathbf{x})\mathbf{I}_y(\mathbf{x}) & \sum_{\mathbf{x}\in P} \mathbf{I}_y^2(\mathbf{x}) \end{pmatrix} \tag{17}$$

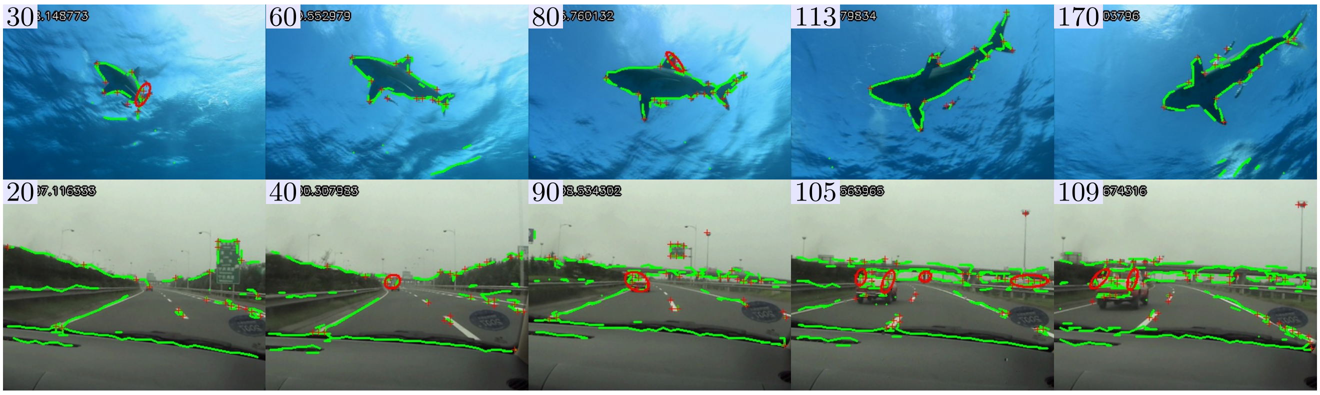

Fig. 14 Experiments on longer sequences.

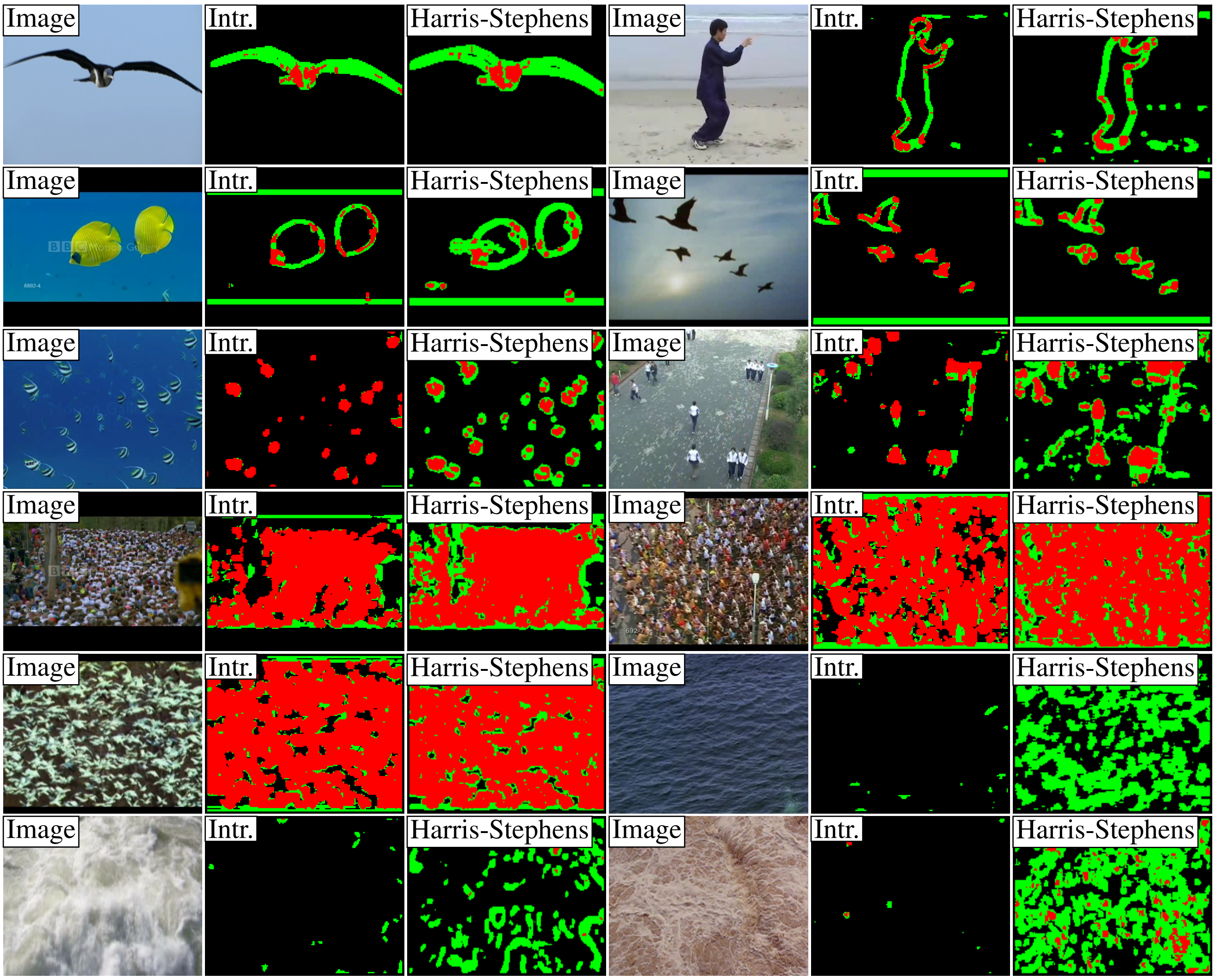

Fig. 15 Comparison with trimaps of the Harris-Stephens detector. For objects at large scales (Row 1∼2), or smaller scales without similar distractions nearby (Row 3), the results of Harris-Stephens and intrackability are similar. For objects at small scales with very similar distractions nearby (Row 4 and left of Row 5), Harris-Stephens is too optimistic. For the textured videos (right of Row 5 and Row 6), Harris-Stephens fails to tell textures from edges.

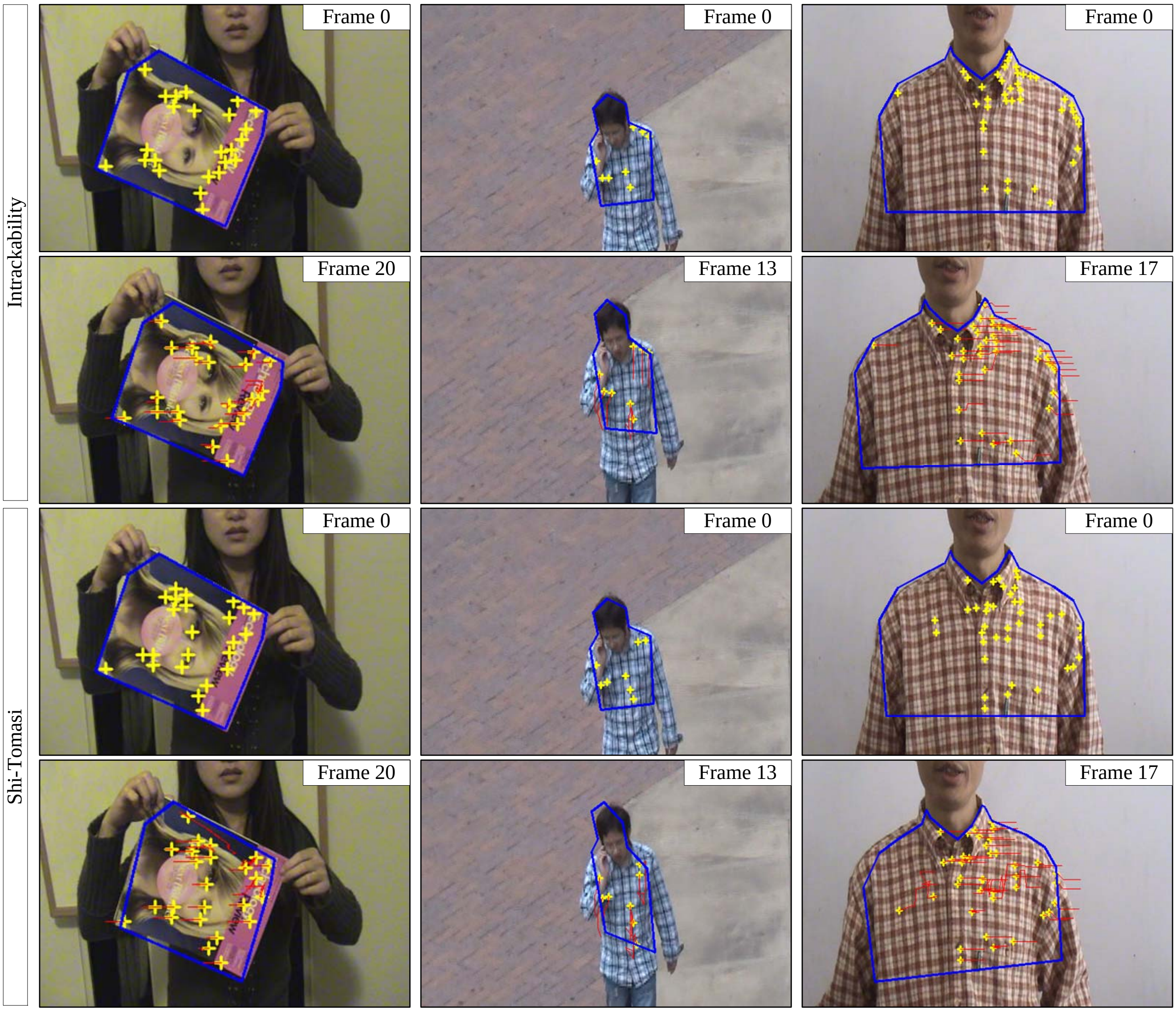

Fig. 16 Tracking comparison: In the first column, the intrackability measure tracks slightly better than Shi-Tomasi measure. In the second and third columns, the intrackability measure can distinguish subtle trackable points from the clothes, but Shi-Tomasi measure selects more repetitive feature points and makes more mismatches across frames.

Let $\lambda_{\max} \geq \lambda_{\min}$ be the two eigen-values of Σ^{-1}, then the local intrackability is

$$\mathcal{H}\{\mathbf{u}|\mathbf{I}, \mathbf{I}'\} = 1 + \log 2\pi + \frac{1}{2}\det(\Sigma),$$
$$= 1 + \log 2\pi - \frac{1}{2}\log \lambda_{\max}\lambda_{\min}.$$

Therefore, large eigen-values leads to lower intrackability and thus to better points to track. In the projected direction $u_\perp$, we drop the dimension that has lower eigen-value, and the intrackability of an oriented line is

$$\mathcal{H}\{u_\perp|\mathbf{I}, \mathbf{I}'\} = \frac{1}{2} + \frac{1}{2}\log 2\pi - \frac{1}{2}\log \lambda_{\max}$$

In comparison, (Shi and Tomasi, 1994) used $\lambda_{\min}$ as a texturedness measure. Larger $\lambda_{\min}$ means higher intensity contrast in the patch and thus a better point to track.

We can see that the differences between intrackability and the Shi-Tomasi measure are

1. Shi-Tomasi uses Taylor expansion as an approximation of a local image patch. This assumes that the image is continuous and may be violated in textured motion.
2. $\lambda_{\min}$ is used instead of $\log \lambda_{\max}\lambda_{\min}$ measure.

It is worth noting that the Shi-Tomasi texturedness measure is most effective in a video regime corresponding to the rightmost extreme in Fig. 5 (bird flock) and Fig. 10 (marathon) where the objects are dense and still distinguishable from the surroundings. In our pursued hybrid representations, most trackable points are selected in this regime in Fig. 13 (row 3).

We compare with (Shi and Tomasi, 1994) in selecting good features to track in frame-to-frame tracking. The Shi-Tomasi criterion measures texturedness in a single image patch of 5×5 pixels, in contrast our intrackability is computed between frames in a $[-12, 12]^2$ displacement range and thus encompasses a larger neighborhood. As Fig. 16 illustrates, we manually initialize a polygonal region for the object of interest, then trackable points are pursued in the region and tracked across frames by finding the best SSD matches. After point-wise matching, an affine transformation is fit to obtain a current polygon for the object region. For an object with no self-similar feature, our results are similar to or slightly better than the Shi-Tomasi measure, see the first column in Fig. 16. But for objects with many self-similar features, the Shi-Tomasi measure will be misguided to choose these self-similar ones, which often results in mismatches between frames. In Figure 16, the second and third column show that the intrackability measure can distinguish the more informative points on collars, shoulders, buttons and pockets in most places, but the Shi-Tomasi measure fails to do so in more places.

To make quantitative comparison of the performances, we annotate the ground truth of the vertices of outer polygons for the three sequences in Fig. 16 and measure the average errors of all vertices over time. Let $\mathbf{x}_{i,t}$ be the ground truth of the position of the i-th vertex in frame t, $\hat{\mathbf{x}}_{i,t}$ be its estimated value by a tracking algorithm, and M be the number of vertices. The tracking error of frame t is defined as

$$\text{Error}_t = \frac{1}{M} \sum_i \|\mathbf{x}_{i,t} - \hat{\mathbf{x}}_{i,t}\| \tag{18}$$

The resultant error curves are shown in Fig. 17.

Harris-Stephens R score (Harris and Stephens, 1988) is also based on the matrix in Eq. (17). It is defined as $R = \det(\Sigma^{-1}) - k\text{trace}(\Sigma^{-1})^2$, which is equivalent to $R = \lambda_{\min} * \lambda_{\max} - k(\lambda_{\min} + \lambda_{\max})^2$, where k is a small weight. It is clear that our intrackability measure $\log(\lambda_{\min} * \lambda_{\max})$ is the log of an upper bound to R score.

5.2 Intrackability and the condition number

(Fan et al, 2006) proposed to use the conditional number of a matrix as an uncertainty measure in tracking a kernel. Unlike point tracking, a kernel tracking uses a histogram feature in a larger scope. Let $\mathbf{h}_0$ be the histogram as a model of the target. In the next frame, mean-shift is used to find the optimal motion vector $\mathbf{u}$ of the target, starting from a predicted position. Let $\mathbf{h}_1$ be the histogram at the predicted position, Fan et al. (Fan et al, 2006) began with the linearized kernel tracking equation system

$$\mathbf{M}\mathbf{u} = \sqrt{\mathbf{h}_0} - \sqrt{\mathbf{h}_1} \tag{19}$$

where $\mathbf{M} = (\mathbf{d}_1, \cdots, \mathbf{d}_m)^T$ is a matrix composed of centers of mass of all color bins and $\mathbf{d}_j$ is the j-th mass center. Let $\mathbf{A} = \mathbf{M}^T\mathbf{M}$ be the matrix with two eigenvalues $\lambda_{\max}$ and $\lambda_{\min}$. The condition number of $\mathbf{A}$ is $\lambda_{\max}/\lambda_{\min} \geqslant 1$. Small condition number will result in a stable solution to Eq 19 and thus a better kernel to track.

To compare with this measure, we rewrite the local posterior probability for the velocity $\mathbf{u}$ according to this setup,

$$p(\mathbf{u}|\mathbf{h}_0, \mathbf{h}_1) \propto \exp\left\{-\frac{\|\mathbf{M}\mathbf{u} - (\sqrt{\mathbf{h}_0} - \sqrt{\mathbf{h}_1})\|^2}{\text{tr}(\mathbf{A})}\right\}. \tag{20}$$

where the trace $\text{tr}(\mathbf{A}) = \lambda_{\max} + \lambda_{\min}$ is introduced to normalize the histogram differences. This is also a two dimensional Gaussian with covariance matrix

$$\Sigma = \text{tr}(\mathbf{A})\mathbf{A}^{-1}. \tag{21}$$

Therefore, the local intrackability is the entropy of $p(\mathbf{u}|\mathbf{h}_0, \mathbf{h}_1)$.

$$\mathcal{H}\{\mathbf{u}|\mathbf{h}_0, \mathbf{h}_1\} = 1 + \log 2\pi - \log \frac{\sqrt{\lambda_{\max}\lambda_{\min}}}{\lambda_{\max} + \lambda_{\min}} \tag{22}$$

$$= 1 + \log 2\pi - \log \frac{\sqrt{\lambda_{\max}/\lambda_{\min}}}{\lambda_{\max}/\lambda_{\min} + 1}. \tag{23}$$

This is a monotonically increasing function with respect to the condition number $\lambda_{\max}/\lambda_{\min}$ as $\lambda_1/\lambda_2 \geqslant 1$.

In light of the same derivation process, other covariance related measures such as those mentioned in (Zhou et al, 2005) can all be regarded as an intrackability under some Gaussian distribution assumption.

6 Discussion

Despite the vast literature in motion analysis, tracking, and video coding, the connections and transitions between various video representations have not been studied. In this paper, we study the intrackabilities of local image entities (points, lines, patches) as a measure of the inferential uncertainty. Using the histogram of the intrackabilities pooled over the video in space and time as the global video statistics, we map natural video clips in a scatter plot and examine the different regimes. We find two major axes in the plot representing image scaling and change of object density respectively. As a video may contain multiple patterns in different regimes, we develop a model selection criterion based on the intrackability and model complexity to pursue a hybrid representation which integrates four components: trackable points, trackable lines, contours, and textured motion. This criterion guides the transition of representations due to image scaling and change of object density.

In representing generic images, researchers have developed sparse coding models for structured image primitives, such as edges, bars, and corners etc. and texture models based on Markov random fields for stochastic textures which

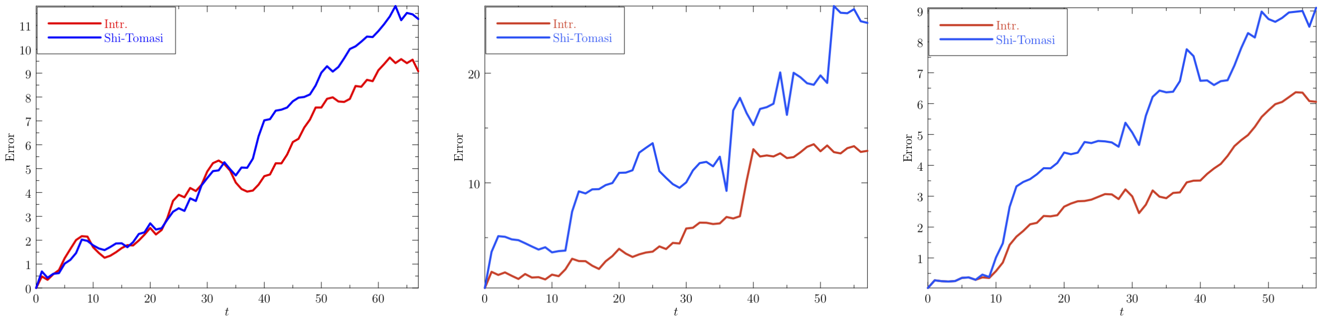

Fig. 17 Quantitative performance comparison — left is the magazine sequence (left column in Fig. 16), middle is the phone-call sequence (middle column in Fig. 16), and right the cloth sequence (right column in Fig. 16).

do not have distinct elements. The integration of these models has led to a primal sketch model conjectured in (Marr et al, 1979). In an ongoing project, we are extending the hybrid representation to a video primal sketch model as a generic video representation for effective coding and for modeling various actions.

Acknowledgement

The authors would like to thank Dr. Yingnian Wu at UCLA for discussions and thank Youdong Zhao at Lotus Hill Institute for assistance. We also thanks Jeffrey Byrne, Qihui Zhu, Weiyu Zhang and Katerina Fragkiadaki for proofreading. This work is supported by NSFC 60832004 and 60875005.

References

Ali S, Shah M (2007) A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In: CVPR

Ali S, Shah M (2008) Floor fields for tracking in high density crowd scenes. In: ECCV

Badrinarayanan V, Perez P, Le Clerc F, Oisel L (2007) On uncertainties, random features and object tracking. In: ICIP

Black MJ, Fleet DJ (2000) Probabilistic detection and tracking of motion boundaries. IJCV 38(3):231–245

Collins R, Liu Y, Leordeanu M (2005) Online selection of discriminative tracking features. PAMI 27(10):1631–1643

Collins RT (2003) Mean-shift blob tracking through scale space. In: CVPR

Comaniciu D, Ramesh V, Meer P (2003) Kernel-based object tracking. PAMI 25(5)

Cong Y, Gong H, Zhu SC, Tang Y (2009) Flow mosaicking: Real-time pedestrian counting without scene-specific learning. In: CVPR, pp 1093–1100

Dreschler L, Nagel HH (1981) Volumetric model and 3D trajectory of a moving car derived from monocular TV frame sequences of a street scene. In: IJCAI, pp 692–697

Fan Z, Yang M, Wu Y, Hua G, Yu T (2006) Effient optimal kernel placement for reliable visual tracking. In: CVPR

Fitzgibbon A (2001) Stochastic ridigity: Image registration for nowhere-static scenes. In: ICCV

Han TX, Ramesh V, Zhu Y, Huang TS (2005) On optimizing template matching via performance characterization. In: ICCV

Harris C, Stephens M (1988) A combined corner and edge detector. In: Proceedings of The Fourth Alvey Vision Conference, Manchester, UK, pp 147–151

Horn B, Schunck B (1981) Determining optical flow. Artificial Intelligence 17:185–203

Kadir T, Brady M (2001) Saliency, scale and image description. IJCV 45(2):83–105

Koenderink JJ (1984) The structure of images. Biological Cybernetics 50:363–370

Kwon J, Lee KM, Park FC (2009) Visual tracking via geometric particle filtering on the affine group with optimal importance function. IEEE Conf on Computer Vision and Pattern Recognition

Li Z, Gong H, Sang N, Zhu G (2007a) Intrackability theory and application. In: SPIE MIPPR

Li Z, Gong H, Zhu SC, Sang N (2007b) Dynamic feature cascade for multiple object tracking with trackability analysis. In: EMMCVPR

Lindeberg T (1993) Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention. IJCV 11(3):283–318

Maccormick J, Blake A (2000) A probabilistic exclusion principle for tracking multiple objects. IJCV 39(1):57–71

Marr D, Poggio T, Ullman S (1979) Bandpass channels, zero-crossings, and early visual information processing. JOSA 69:914–916

Nickels K, Hutchinson S (2002) Estimating uncertainty in ssd-based feature tracking. Image and Vision Computing 20(1):47–68

Pan P, Porikli F, Schonfeld D (2009) Recurrent tracking using multifold consistency. In: IEEE Workshop on VS-PETS

Pylyshyn ZW (2004) Some puzzling findings in multiple object tracking (MOT): I. tracking without keeping track of object identities. Visual Cognition 11(7):801–822

Pylyshyn ZW (2006) Some puzzling findings in multiple object tracking (mot): II. inhibition of moving nontargets. Visual Cognition 14(2):175–198

Pylyshyn ZW, Vidal Annan J (2006) Dynamics of target selection in multiple object tracking (mot). Spatial Vision 19(6):485–504

Ross DA, Lim J, Lin RS, Yang MH (2008) Incremental learning for robust visual tracking. Int'l Journal of Computer Vision 77:125–141

Sato K, Aggarwal JK (2004) Temporal spatio-velocity transform and its application to tracking and interaction. CVIU 96:100–128

Segviĉ S, Remazeilles A, Chaumette F (2006) Enhancing the point feature tracker by adaptive modelling of the feature support. In: ECCV

Serby D, Koller-Meier S, Gool LV (2004) Probabilistic object tracking using multiple features. In: ICPR, pp 184–187

Shi J, Tomasi C (1994) Good features to track. In: CVPR

Soatto S, Doretto G, Wu Y (2001) Dynamic textures. In: ICCV

Srivastava A, Lee A, Simoncelli E, Zhu S (2003) On advances in statistical modeling of natural images. J of Math Imaging and Vision 18(1):17–33

Szummer M, Picard RW (1996) Temporal texture modeling. In: ICIP

Tommasini T, Fusiello A, Trucco E, Roberto V (1998) Making good features track better. In: CVPR

Veenman C, Reinders M, Backer E (2001) Resolving motion correspondence for densely moving points. PAMI 23:54–72

Wang Y, Zhu S (2008) Perceptual scale space and its applications. Int'l Journal of Computer Vision 80(1):143–165

Wang Y, Zhu SC (2003) Modeling textured motion : Particle, wave and sketch. In: ICCV, pp 213–220

Wang Y, Bahrami S, Zhu SC (2005) Perceptual scale space and its applications. In: ICCV, pp 58–65

Witkin A (1983) Scale space filtering. In: Intl Joint Conf. on AI, Kaufman, Palo Alto

Wu Y, Zhu S, Guo C (2008) From information scaling of natural images to regimes of statistical models. Quarterly of Applied Mathematics 66(1):81–122

Yilmaz A, Javed O, Shah M (2006) Object tracking: A survey. ACM Computing Survey 38(4):13

Zhou XS, Comaniciu D, Gupta A (2005) An information fusion framework for robust shape tracking. PAMI 27(1):115–123