A STUDY OF ACTOR AND ACTION SEMANTIC RETENTION IN VIDEO SUPERVOXEL SEGMENTATION

CHENLIANG XU[*‡] and RICHARD F. DOELL[†‡]

Department of Compute Science and Engineering
SUNY at Buffalo, Buffalo, NY 14260, USA
[*]*chenlian@buffalo.edu*
[†]*rfdoell@buffalo.edu*
[‡]*Contributed equally to this work.*

STEPHEN JOSÉ HANSON[§] and CATHERINE HANSON[¶]

Department of Psychology and Rutgers Brain Imaging Center
Rutgers University, Newark, NJ 07102, USA
[§]*jose@psychology.rutgers.edu*
[¶]*cat@psychology.rutgers.edu*

JASON J. CORSO

Department of Compute Science and Engineering
SUNY at Buffalo, Buffalo, NY 14260, USA
jcorso@buffalo.edu

Existing methods in the semantic computer vision community seem unable to deal with the explosion and richness of modern, open-source and social video content. Although sophisticated methods such as object detection or bag-of-words models have been well studied, they typically operate on low level features and ultimately suffer from either scalability issues or a lack of semantic meaning. On the other hand, video supervoxel segmentation has recently been established and applied to large scale data processing, which potentially serves as an intermediate representation to high level video semantic extraction. The supervoxels are rich decompositions of the video content: they capture object shape and motion well. However, it is not yet known if the supervoxel segmentation retains the semantics of the underlying video content. In this paper, we conduct a systematic study of how well the actor and action semantics are retained in video supervoxel segmentation. Our study has human observers watching supervoxel segmentation videos and trying to discriminate both actor (human or animal) and action (one of eight everyday actions). We gather and analyze a large set of 640 human perceptions over 96 videos in 3 different supervoxel scales. Furthermore, we design a feature defined on supervoxel segmentation, called supervoxel shape context, which is inspired by the higher order processes in human perception. We conduct actor and action classification experiments with this new feature and compare to various traditional video features. Our ultimate findings suggest that a significant amount of semantics have been well retained in the video supervoxel segmentation and can be used for further video analysis.

Keywords: Semantic retention; computer vision; video supervoxel segmentation; action recognition.

2   *Chenliang Xu, Richard F. Doell, Stephen José Hanson, Catherine Hanson, and Jason J. Corso*

1. Introduction

We are drowning in video content—YouTube, for example, receives 72 hours of video uploaded every minute. In many applications, there is so much video content that a sufficient supply of human observers to manually tag or annotate the videos is unavailable. Furthermore, it is widely known that the titles and tags on the social media sites like Flickr and YouTube are noisy and semantically ambiguous [1]. Automatic methods are needed to index and catalog the salient content in these videos in a manner that retains the semantics of the content to facilitate subsequent search and ontology learning applications.

However, despite recent advances in computer vision, such as the deformable parts model for object detection [2], the scalability as the semantic space growing remains a challenge. For example, the state of the art methods on the ImageNet Large Scale Visual Recognition Challenge [3] have accuracies near 20% [4] and a recent work achieves a mean average precision of 16% on a 100,000 class detection problem [5], which is the largest such multi-class detection model to date. To compound this difficulty, these advances are primarily on images and not videos. Methods in video analysis, in contrast, still primarily rely on low-level features [6], such as space-time interest points [7], histograms of oriented 3D gradients [8], or dense trajectories [9]. These low-level methods cannot guarantee retention of any semantic information and subsequent indices likewise may struggle to mirror human visual semantics. More recently, a high-level video feature, called Action Bank [10], explicitly represents a video by embedding it in a space spanned by a set, or bank, of different individual actions. Although some semantic transfer is plausible with Action Bank, it is computationally intensive and struggles to scale with the size of the semantic space; it is also limited in its ability to deduce viewpoint invariant actions.

In contrast, segmentation of the video into spatiotemporal regions with homogeneous character, called *supervoxels*, has a strong potential to overcome these limitations. Supervoxels are significantly fewer in number than the original pixels and frequently surpass the low-level features as well, and yet they capture strong features such as motion and shape, which can be used in retention of the semantics of the underlying video content. Figure 1 shows an example supervoxel segmentation generated by the streaming hierarchical supervoxel method [11]. The individual supervoxels are denoted as same-colored regions over time in a video. Furthermore, results in the visual psychophysics literature demonstrate that higher order processes in human perception rely on shape [12] and boundaries [13, 14, 15]. For instance, during image/video understanding, object boundaries are interpolated to account for occlusions [13] and deblurred during motion [15]. However, the degree to which the human semantics of the video content are retained in the final segmentation is unclear. Ultimately, a better understanding of semantic retention in video supervoxel segmentation could pave the way for the future of automatic video understanding methods.
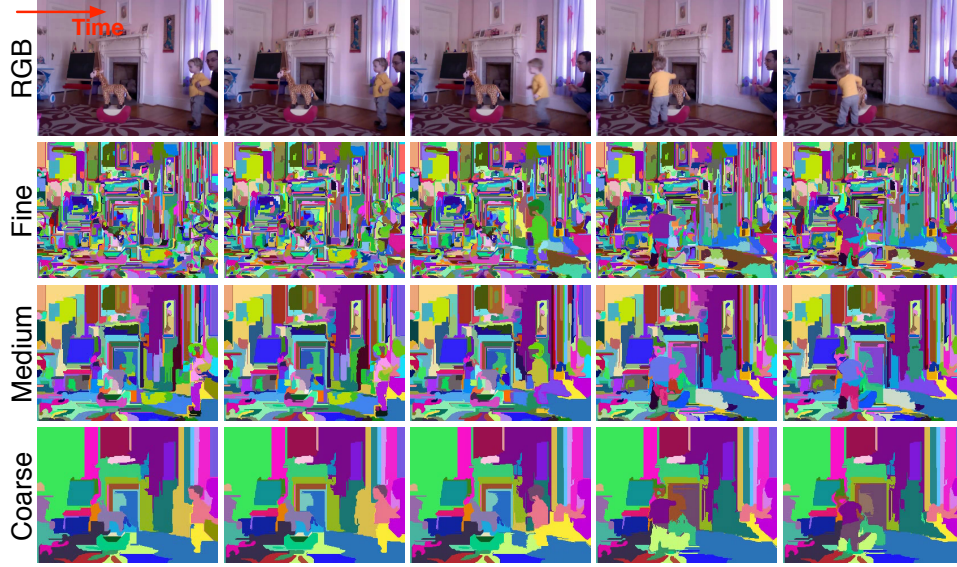
Fig. 1. Example output of the streaming hierarchical supervoxel method. Columns from left to right are frames uniformly sampled from a video. Rows from top to bottom are: the original RGB video, the fine segmentation (low level in the hierarchy), the medium segmentation (middle level in the hierarchy), and the coarse segmentation (high level in the hierarchy).

To that end, we conduct a systematic study of how well the actor and action semantics in moving videos are retained through various supervoxel segmentations. Concretely, we pose and answer the following five questions:

(1) Do the segmentation hierarchies retain enough information for the human perceiver to discriminate actor and action?
(2) How does the semantic retention vary with density of the supervoxels?
(3) How does the semantic retention vary with actor?
(4) How does the semantic retention vary with static versus moving background?
(5) How does response time vary with action?

A preliminary version of our study appeared in [16], in which we present novice human observers with supervoxel segmentation videos (i.e., not RGB color videos but supervoxel segmentation videos of RGB color videos) and ask them to, as quickly as possible, determine the actor (*human* or *animal*) and the action (one of eight everyday actions such as *walking* and *eating*). The system records these human perceptions as well as the response time and then scores whether or not they match the ground truth perceptions; if so, then we consider that the semantics of the actor/action have been retained in the supervoxel segmentation. We systematically conduct the study with a cohort of 20 participants and 96 videos. Ultimately, the human perception results indicate that a significant amount of semantics have been retained in the supervoxel segmentation.

In addition, we conduct machine recognition experiments on actor and action with a feature defined on the supervoxel segmentation, called *supervoxel shape context*, and compare it with various video features, such as dense trajectories [9] and action bank [10]. The supervoxel shape context captures the important shape information of supervoxels, which is inspired by the shape context on still images [17]. Our experimental results suggest that the underlying semantics in supervoxel segmentation can be well used in machine recognition to achieve competitive results, and the overall machine recognition of actor and action seems to follow the same trend as that of human perception but more work needs to be done to get the machine recognition models up to par with the humans in terms of recognition performance.

The remainder of the paper is organized as follows. Section 2 provides the background on video supervoxel segmentation. Section 3 describes the details of the data set acquisition as well as the experiment setup for both human perception and machine recognition using supervoxel segmentations. Section 4 dicusses the results and our analysis of the underlying semantics and recognition experiment. Finally, Section 5 concludes our findings.

2. Video Supervoxel Segmentation

2.1. *Supervoxel Definition, Methods and Applications*

Supervoxel Definition. Perceptual grouping of pixels into roughly homogeneous and more computationally manageable regions, called *superpixels*, has become a staple of early image processing [18, 19]. Supervoxels are the video analog to the image superpixels. Recently, supervoxel segmentation has risen as a plausible first step in early video processing [20, 21, 22, 11]. Consider the following general mathematical definition of supervoxels, as given in our early work [22], where a video is treated as a 3D volume composed by frames with time as the third dimension. We build a 3D lattice Λ^3 composed by voxels (pixels in a video) with edges linking neighbors in space and time, and define a *supervoxel s* as a subset of the lattice $s \subset \Lambda^3$ such that the union of all supervoxels comprises the lattice and they are pairwise disjoint: $\bigcup_i s_i = \Lambda^3 \wedge s_i \bigcap s_j = \varnothing \ \forall i, j$ pairs.

Although the lattice Λ^3 itself is indeed a supervoxel segmentation, it is far from a so-called *good* one [22]. Typical algorithms seek to enforce principles of spatiotemporal grouping—proximity, similarity and continuation—from classical Gestalt theory [23, 24], boundary preservation, and parsimony. From the perspective of machine vision, the main rationale behind supervoxel oversegmentation is two fold: (1) voxels are not natural elements but merely a consequence of the discrete sampling of the digital videos and (2) the number of voxels is very high, making many sophisticated methods computationally infeasible. Therefore, supervoxels serve as an important data representation of a video, such that various image/video features may be computed on the supervoxels, including color histograms, textons, etc.

Supervoxel Methods. Methods for generating video supervoxel segmentation are

thoroughly evaluated in our proposed LIBSVX[a] benchmark [22, 11, 25], which implements a suite of five supervoxel algorithms and tests them in a set of evaluation metrics with three video datasets. The methods in the library include segmentation by weighted aggregation [26, 27], graph-based [28], hierarchical graph-based [21], mean shift [29], and Nyström normalized cuts [30]. Ultimately, it was determined that the two hierarchical agglomerative methods [26, 27, 21] perform best overall due to the way in which multiscale region similarity was reevaluated as the hierarchy was generated. We note that later methods based on temporal extension of superpixels [31, 32, 33, 34] and hierarchical trajectory merging [35] also achieve competitive performance with respect to the library's benchmark. In this paper, we use the streaming version [11] of hierarchical graph-based algorithm [21] to generate the supervoxels due to its good performance and the ability to handle long videos.

Supervoxel Applications. Despite the richness in supervoxel methods and the well developed evaluation benchmark, supervoxels are still treated as merely groupings of video based features for object and region labeling problem. Tighe and Lazebnik [36] use supervoxels for region labeling in videos, where standard video features such as color histograms and optical flow [37] are computed. Raza et al. [38] build geometric context classifiers on supervoxels by exploring motion based features and appearance features. Tang et al. [39] propose a weakly supervised algorithm to tackle the large-scale video object labeling problem. However, the ability of using supervoxels for actor and action recognition remains as an unexplored research problem.

2.2. *Streaming Hierarchical Supervoxel Method*

We use the state of the art streaming hierarchical supervoxel method [11] to generate a supervoxel segmentation hierarchy $\mathcal{S} = \{S^1, S^2, \ldots, S^H\}$ of an input video $\mathcal{V}$, where $S^h = \{s_i^h\}, h \in \{1, 2, \ldots, H\}$ is the supervoxel segmentation at level h in the hierarchy. The levels of a supervoxel hierarchy define the merging stages of supervoxels. In our case, the size of supervoxel s grows from fine to coarse when varying h from 1 to H. The method obtains the hierarchical segmentation result $\mathcal{S}$ by minimizing:

$$\mathcal{S}^* = \underset{\mathcal{S}}{\operatorname{argmin}} \, E(\mathcal{S}|\mathcal{V}) \ , \tag{1}$$

where the objective criterion $E(\cdot|\cdot)$ is defined by the minimum spanning tree method in [28]. For example, for the hth level in the hierarchy, the objective criterion is defined as:

$$E(S^h|\mathcal{V}) = \tau \sum_{s \in S^h} \sum_{e \in \mathrm{MST}(s)} w(e) + \sum_{s,t \in S^h} \min_{e \in <s,t>} w(e) \ , \tag{2}$$

where, $\mathrm{MST}(s)$ denotes the minimum spanning tree (of voxels or supervoxels from the previous fine level in the hierarchy) in the supervoxel s, e is the edge defined

---

[a]`http://www.cse.buffalo.edu/~jcorso/r/supervoxels/`

6   *Chenliang Xu, Richard F. Doell, Stephen José Hanson, Catherine Hanson, and Jason J. Corso*
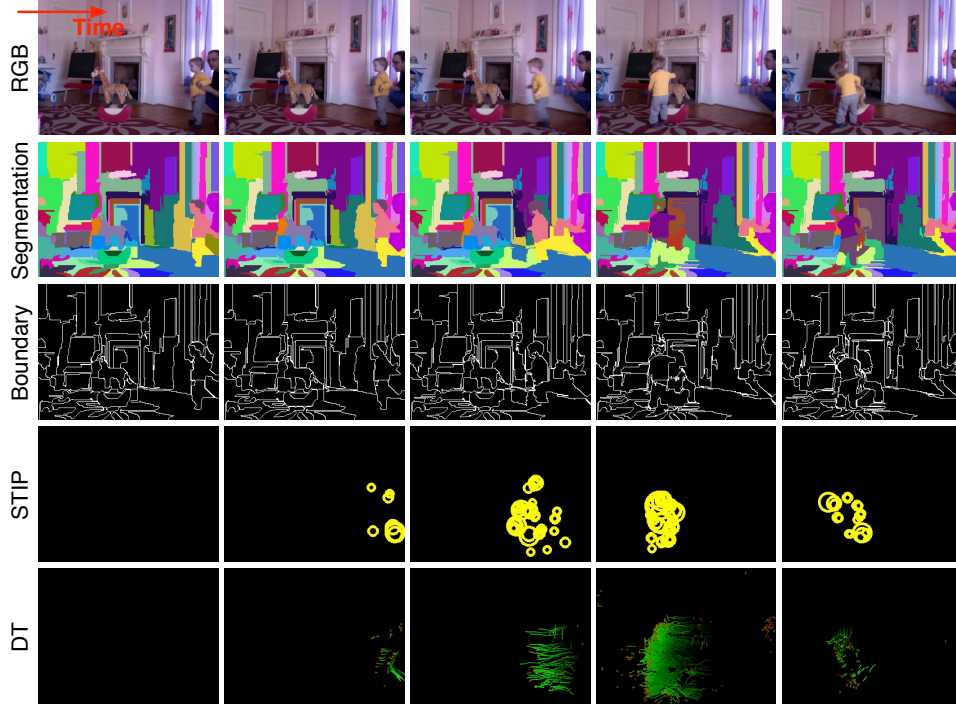

Fig. 2. A comparison of different video feature representations. From top to bottom rows are: the RGB video, the supervoxel segmentation, extracted boundaries of supervoxel segmentation, space-time interest points (STIP), and dense trajectories (DT).

by the 3D lattice Λ^3, $w(e)$ is the edge weight, and τ is a parameter that balances the two parts. The edge weight $w(e)$ captures the color differences of voxels. By minimizing Eq. (1), the algorithm ultimately outputs a supervoxel segmentation hierarchy of the original input RGB video.

Figure 1 shows a hierarchical supervoxel segmentation produced by [11]. The segmentations from top to bottom rows are sampled from low, middle, and high levels in a supervoxel segmentation hierarchy, where each have fine, medium and coarse segments respectively. Each supervoxel has a unique color and we randomly color the output supervoxels in one level with the constraint that the same color is not shared by different supervoxels. In general, we allow reuse of colors in different levels in the segmentation hierarchy, since they are not used in a single run of human perception experiment in this work.

2.3. *Supervoxels: Rich Decompositions of RGB Videos*

Considering the example in Figure 1, we observe that the hierarchy of the supervoxel segmentation captures different levels of semantics of the original RGB video. For example, one tends to recognize the humans easier from coarser levels in the

hierarchy, since they are captured by fewer supervoxels; however, the coarser levels lose the detailed content in the video, such as the woman in the painting hanging on the wall, which is still captured at the medium level.

Compared with other features, such as space-time interest points (STIP) [7] and dense trajectories (DT) [9], which are frequently used in video analysis [6], the supervoxel segmentation seems to retain more semantics of the RGB video (in this paper we seek to quantify how many of these semantics are retained for one set of actors and actions). Figure 2 shows a visual comparison among those features. STIP and DT use the sampled points and trajectories as the data representation—this is not the full STIP or DT feature descriptor representation, which also measures other information, such as gradient.

By only watching the videos of STIP and DT, as shown in the bottom two rows of Figure 2, it seems unlikely that humans could recover the content of a video, especially when there is little motion in a video. On the other hand, one can easily recover the content of a video by watching the supervoxel segmentation video, likely due to the fact that the supervoxel segmentation retains the shape of the objects (boundaries of the supervoxel segmentation are also shown in the third row of Figure 2). Zitnick and Parikh [40] show that the segmentation boundaries are in general better than classical edge detection methods, such as those generated by the Canny edge detector[41], for automatic image understanding, and they perform as well as humans using low-level cues.

The precise goals of this paper are to explore exactly how much semantic content, specifically the actor (human or animal) and the action, is retained in the supervoxel segmentation and can they be used for machine recognition for actors and actions. We describe the experiment and results of supervoxel human perception and machine recognition in the next two sections.

3. Experiment Setup

We have set up a systematic experiment to study actor and action semantic retention in the video supervoxel segmentation. By actor we simply mean *human* or *animal*. For action, we include a set of eight actions: *climbing*, *crawling*, *eating*, *flying*, *jumping*, *running*, *spinning* and *walking*. We have gathered a complete set of videos (Section 3.1) and processed them through the segmentation algorithm (Section 2). Then we use the data to conduct human perception experiments (Section 3.2) as well as machine recognition experiments (Section 3.3). Finally, we analyze the aggregate results over the full data cohort and quantify the retention of semantics (Section 4).

3.1. *Data Set*

We have collected a data set with two kinds of actors (*humans* and *animals*) performing eight different actions: *climbing*, *crawling*, *eating*, *flying*, *jumping*, *running*, *spinning* and *walking*. We only include animals that frequently appear in (North

8   *Chenliang Xu, Richard F. Doell, Stephen José Hanson, Catherine Hanson, and Jason J. Corso*

Fig. 3. A snapshot of the RGB videos in our data set. The actors in the top two rows are humans and in the bottom two rows are animals. The data set consists of two kinds of actors, eight actions and two types of background settings, resulting in a total of 32 videos.

American) daily life, such as dogs, cats, birds, squirrels and horses. The backgrounds of the videos fall into two categories: static (relatively static objects such as ground and buildings with little camera changes) and moving (moving objects in the background, such as in a traffic or dramatic camera changes). A complete RGB video data set consists of 32 videos in total (2 actors × 8 actions × 2 background types = 32). Figure 3 shows a snapshot of the RGB videos we collected.

Each video in the data set is about four seconds long and the actor starts the action immediately after the video plays. We, however, show the videos at half-frame-rate when conducting the experiment to allow ample response time for the human participants. We have attempted to exclude those videos that have ambiguity with either the actors or the actions, and only use videos that have a major actor performing one single action. For example, a disqualified human jumping video usually contains running before jumping. But, some ambiguity remains due to the general complexity of dynamic video. The data set used in this paper is a complete data set having a single video for each actor, action and background type tuple. All videos were downloaded from public "wild" repositories, such as YouTube.

For each of the RGB videos, we use the method described in Section 2.2 to obtain a supervoxel segmentation hierarchy. We first use ffmpeg to resize the videos to 320x240 maintaining the original aspect ratio. Then, we run the *gbh˙stream* program (LIBSVX version 2.0) with the following parameters: `c: 0.2`, `c_reg: 10`, `min: 20`, `sigma: 0.4`, `range: 10`, `hie_num: 30`. Note that `c` and `c_reg` map to τ in Eq. (2) (`c` is used at the first level and `c_reg` is used at all other hierarchy levels).

We sample three different levels (fine: 8th level, medium: 16th level, coarse: 24th level) from the hierarchy, similar as in Figure 1. Therefore, the full set of data we used to run the semantic retention experiment is the 96 supervoxel segmentation videos of three different levels. Note that the audio is disabled, so that the par-

ticipants in human perception expeiment only have the vision perception of the supervoxel segmentation videos (and never the RGB videos).

3.2. *Human Perception Experiment*

The purpose of the human perception experiment is to measure exactly how much semantic content is retained in the supervoxel segmentation. In the perception experiment, we show the supervoxel segmentation videos to human observers and request them to make a forced-choice selection of actor and action.

3.2.1. *Data Split and Study Cohort*

We create a threeway split of the 96 videos into three sets: alpha, beta and gamma. Since each of the original 32 videos is represented in three levels of the hierarchy, it is imperative to make the threeway split and thereby avoid one participant seeing the same video twice but on two different supervoxel hierarchy levels. So, each of alpha, beta and gamma have the full 32 videos, but on different hierarchy levels and uniformly varying over levels in each of the three splits. Based on the ordering in the database, alpha will start with one level (say coarse) in the hierarchy, then beta will have the next (medium) and gamma the third (fine) for one original RGB video. For the next original RGB video, it will rotate, so that alpha has the next level (now medium), beta the next (fine) and gamma will wrap around to the first (coarse) again. This repeats through all 96 supervoxel videos. Before the videos are shown to the participant, the order of the videos is shuffled, so that the participant cannot deduce the contents based on an ordering of the videos (like human human human ... animal animal animal).

Our study cohort is 20 college-age participants. To ensure generality, we exclude those students who are studying video segmentation (and hence may have already developed an *eye* for semantic content in supervoxel segmentations). Each participant is shown one split of the videos (alpha, beta or gamma). And each participant sees a given video only once. Participants never see the input RGB videos.

3.2.2. *User Interface and Instructions*

We developed a web-based user interface to collect human responses. Figure 4 shows a snapshot of it. The left part of the participant's screen is the supervoxel segmentation video and the right part of the participant's screen comprises two sets of buttons that allow the user to choose either human or animal as the actor, and to choose one of the eight actions. The participant has the option to choose *unknown* (the option "Don't know act or actor" is shown in the center of the select action area in Figure 4). Such an *unknown* option prevents the participant from random selection.

Initially, when the participant starts the experiment, the left part of the screen is blank and buttons on the right side are deactivated (grayed out); once the next video

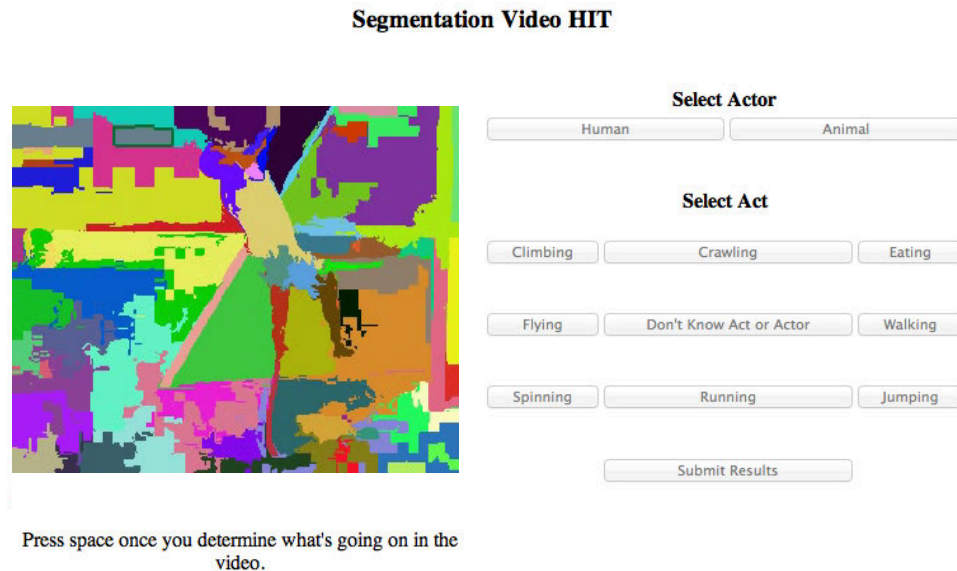10 *Chenliang Xu, Richard F. Doell, Stephen José Hanson, Catherine Hanson, and Jason J. Corso*

Fig. 4. A snapshot of the user interface for the human perception experiment. The left part of screen is the supervoxel segmentation video and the right part of screen is a set of selection buttons. Users are instructed that time to make a selection is recorded and important for the human perception experiment.

in their split is downloaded locally, it prompts the user with a "ready" message. As soon as the participant presses the *space* key, it starts to show the supervoxel segmentation video and the interface triggers a timer that records the response time of the participant. The participant is required to respond by pressing the *space* key again as soon as he or she captured enough information to reach a decision (i.e., knows the actor and action in the supervoxel video). The amount of time between these two *space* key pressing is recorded as one's response time. After this second *space* key is hit, the buttons on the right side are activated and ready for the participant to select them. The participant can only watch the video once, which means once the video reaches the end, the participant is forced to make a decision (or choose unknown) without the option to watch it again. In this case, the whole video time is recorded as one's response time. This process is repeated for each video in the split (alpha, beta or gamma) until the end.

Before a participant begins, s/he is instructed briefly as to the nature of the experiment (trying to recognize the actor and action in a supervoxel video) and walked through the user interface. They are instructed that time is recorded and important; they should hence stop the video as soon as they know the answer. They are not shown any example supervoxel video before the experiment starts.

3.3. *Machine Recognition Experiment*

We design experiments to explore whether machines can use the features extracted from the same supervoxel segmentation videos that we have shown to human observers (no original pixel RGB videos) to perform the actor and action recognition task. Note that our data assumption is different than the labeling problem in Section 2.1, where both original video and supervoxel segmentation are available and supervoxels are merely groupings of video features. We instead extract features based on supervoxel shape, which is directly observable by the humans (see supervoxel boundary in Figure 2). In our machine recognition experiment, we train a set of classifiers based on the supervoxel shape feature to discriminate actors and actions, then compare the video classification results to state of the art action recognition features, such as dense trajectories [9] and action bank [10], which require the availability of pixel RGB videos. Again, we conduct experiments on supervoxel segmentations at three different hierarchy levels.

3.3.1. *Supervoxel Shape Context (SSC)*

Supervoxel segmentation contains important shape and boundary information that is directly observable by humans, and the results in the visual psychophysics literature [12, 13, 14, 15] further demonstrate that such information helps higher order processes in human perception. Therefore, it is our hypothesis that it is possible for machines to mimic the ability of human observers to discern actor and action by using shape related features extracted on supervoxel segmentation videos, even without the availability of pixel RGB videos. Belongie et al. [17] demonstrate the success of *shape context* feature on shape maching and object recognition. Inspired by the original shape context work, we adapt the descriptor to supervoxel segmentation video for feature extraction. The shape context descriptor at a *reference point* captures the distribution of the (supervoxel) edges relative to it, thus offering a globally discriminative characterization.

The original shape context descriptor is intended for static images. We therefore apply the following modifications to extend it to capture the temporal aspect of supervoxel segmentation videos. First of all, we calculate the center of mass of the optical flow [37] for each frame in a video. This point is indicative of the greatest amount of action occurring in a frame and considered as an approximation of the actor center with respect to our single actor videos. Note that other methods to obtain this point are possible, such as using the center of STIP [7] or DT [9] (see Figure 2). We choose optical flow due to the fast computation and parsimony—it is used when generating supervoxel segmentations [21]. Thus, the set of points is suitable as reference points to calculate the supervoxel shape context, which reduces the complexity required for directly determining the location in a spatiotemporal 3D video volume. The supervoxel shape context as used in this paper are the log-polar histograms of supervoxel edges calculated at the reference point on a per-frame basis. We then use five radial and 12 angular bins for the histogram quantization.

Therefore, the supervoxel shape context in a video capture both spatial information by counting the locations and density of the supervoxel edges nearby the center of motion and also temporal information pertinent to action by utilizing the optical flow as an indicator of which points are most salient.

3.3.2. *Other Video-Based Features*

We include the state of the art video-based features for action recognition in the comparison. The span of the features include both low-level and high-level in terms of underlying semantics. Recall we visually compared the supervoxel segmentation with the other video features in Section 2.3.

Low-level Video Feature. The low-level video features we include are dense trajectories [9] and HOG3D [8]. Dense trajectories are calculated from optical flow in a video and the descriptors on the trajectories capture the shape (point coordinates), appearance (histograms of oriented gradients) and motion (histograms of optical flow), as well as the differential in motion boundry (motion boundry histograms). A dense representation of trajectories guarantees a good coverage of foreground motion as well as of the surrounding context. HOG3D is the video extension of the histogram of oriented gradients based descriptors for static images [42]. It treats videos as spatio-temporal volumes and generalizes the key histogram of gradients concepts to 3D and thus it captures the global shape information in a video. For both of these low-level features, we use a bag-of-words classification scheme [43].

High-level Video Feature. Action bank [10] is a high-level representation of video. It is comprised of many individual action detectors, such as boxing and horse back riding, that are sampled broadly in semantic space as well as viewpoint space. It computes the responses of those individual action detectors at different locations and scales in a video and ultimately transfers the video to a response vector in high-dimensional *action-space*.

Dense trajectories and action bank achieve state-of-the-art action recognition performance on popular computer vision datasets, such as UCF50 [44].

3.3.3. *Classification Setup*

We set up the experiment to classify videos based on actor and action separately. For actor, it is a two class (human or animal) classification task. For action, it is *one versus all* for each of the eight actions. For example, when we train a classifier for action *walking*, we use videos containing walking as positive samples and the rest videos containing the other seven actions as negative samples. Due to the small size of our experiment data set, the classification is done with a *leave-one-out* setup.

For supervoxel shape context (SSC), we again extract the features at different levels in a hierarchy: fine, median and coarse segmentations, which are the same as in human experiment. For dense trajectories (DT) and HOG3D, we use a codebook of 20 words, which is small but necessary due to the data set size and empirically

Table 1. Confusion matrix for actor discrimination.

	un	hu	an
unknown	0	0	0
human	0.11	0.86	0.03
animal	0.17	0.05	0.78

found to outperform other sized codebooks—we also used codebooks of 100 and 1000 words. We note that all the action detectors in action bank are from human-based templates, which limits its performance in actor classification. We refer to [10] for a complete list of action templates.

4. Results and Analysis

Human Perception Responses. The response of a single video by one human participant is defined as: $< actor, action, response\ time >$. In total, we have 640 such supervoxel human perceptions collected (32 videos in each split $\times$ 20 participants). The original RGB videos are used as the ground truth data to measure the *match* of the supervoxel human perceptions. We also measure the response time of the participants for both *matched* perceptions and *unmatched* perceptions. Our analysis in Section 4.1 to Section 4.6 is organized systematically according to five key questions regarding semantic retention.

Machine Recognition Results. Our machine classification setting is different than the human perception response. We put actor and action recognition in two separate tasks. For actor, it is a two class (human or animal) classification task. For action, it is one versus all for each of the eight actions. For example, given all 32 supervoxel segmentation videos in one hierarchy level, we classify them as *running* or *not running*. We analyze the machine recognition results in Section 4.7.

4.1. *Do the segmentation hierarchies retain enough information for the human perceiver to discriminate actor and action?*

4.1.1. *Actor Discrimination*

Table 1 shows a confusion matrix of the actor discrimination. As high as 86% of the human perceptions correctly identify the human actors, 78% for the animal actors, and in average 82% for actors in general. We also note that participants tend to choose the *unknown* option when they are less confident of the supervoxel segmentation. There is only a small portion of unmatched perceptions 3% and 5% that mistake human as animal or vice versa. This is hence strong evidence showing that the supervoxel segmentation indeed has the ability to retain the actor semantics from the original RGB videos. We suspect this binary discrimination performance

14 *Chenliang Xu, Richard F. Doell, Stephen José Hanson, Catherine Hanson, and Jason J. Corso*

Table 2. Confusion matrix for action discrimination.

	un	wl	sp	rn	jm	ea	cl	cr	fl
unknown	0	0	0	0	0	0	0	0	0
walking	0.11	0.57	0.12	0.12	0	0.01	0.01	0.04	0
spinning	0.15	0.06	0.65	0.03	0	0	0.01	0.04	0.06
running	0.01	0.07	0.07	0.79	0.04	0	0	0.01	0
jumping	0.19	0.01	0.04	0.09	0.57	0	0	0.01	0.09
eating	0.19	0	0	0	0	0.76	0.04	0	0.01
climbing	0.06	0.01	0	0	0.03	0	0.90	0	0
crawling	0.20	0.03	0	0.06	0.01	0	0.01	0.69	0
flying	0.19	0.03	0.01	0	0.01	0.01	0.03	0.03	0.70

is so high because the data cohort includes videos with one dominant actor and the human participant is able to localize this actor with the supervoxel motion information and then use the supervoxel shape information to determine human or animal. We suspect the reason why the human perception of animal actors is less than that of human actors is because the animals in the data set vary more broadly in relative location and orientation than the humans do.

4.1.2. *Action Discrimination*

Table 2 shows a confusion matrix of the action discrimination. The top three scoring actions are climbing (90%), running (79%), and eating (76%), while the bottom three ones are walking (57%), jumping (57%), and spinning (65%). On average, 70.4% of human perceptions correctly match the actions. Of the lower performing actions, only walking has been easily confused with the other actions (12% to spinning and 12% to running, which may be due to semantic ambiguity—see the example of the human walking in the spinning wheel in Figure 9); jumping and spinning have more *unknowns* (19% and 15% respectively) rather than being confused with other actions. An interesting point to observe is that running and climbing are perceived *unknown* significantly fewer than the other six actions. We suspect this is due to the dominant unidirectional motion of these two actions. Overall, this is more strong evidence that suggests the supervoxel segmentation can well retain the action semantics from the original RGB videos.
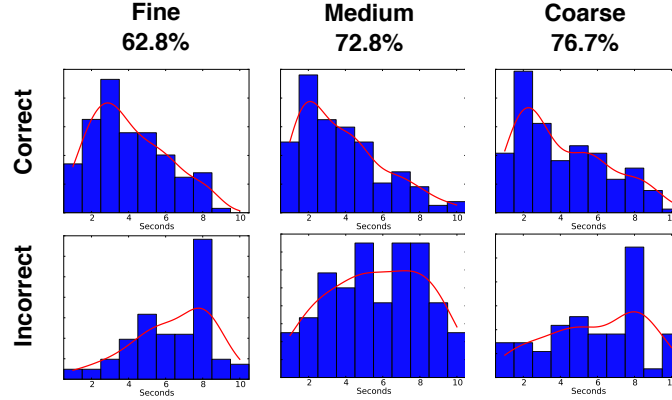
Fig. 5. The performance of supervoxel semantic retention of actor and action on three levels from the supervoxel segmentation hierarchy: fine, medium and coarse. The percentages on top are computed when the supervoxel human perceptions are correctly matched to ground truth ones. The middle and bottom rows are the response time figures when the supervoxel human perceptions are correctly matched and incorrectly matched respectively.

4.2. *How does the semantic retention vary with density of the supervoxels?*

Following the discussion of supervoxel hierarchy in Section 2.3, we seek to understand how the supervoxel size influences retention of action and actor semantics. Recall that we sampled three levels from the supervoxel hierarchy to obtain fine, medium and coarse level supervoxel segmentations. Figure 5 shows the overall performance of the supervoxel human perceptions on different levels. The percentage of correctly matched human perceptions increases when the size of supervoxels grows, suggesting that coarse segmentations more readily retain the semantics of the action and that even coarser segmentations could perform better (i.e., the perfect segmentation of the actor performing the action would likely perform best). A second plausible explanation is that for actor and action discrimination the finer details in the other levels are unlikely to be needed.

We also show study of the response time in Figure 5. Here, we plot the density of responses (horizontal axis is time, at half-frame-rate; vertical axis is density). The blue bars are a simple histogram and the red curve is a Gaussian kernel density estimate. For correct action matches, the response distributions are nearly equivalent, and are heavily weighted toward the shorter end of the plot, indicating that if the participant knows the answer then typically knows it quickly. However, for the incorrect matches, we see different patterns, the fine videos are peaked at about eight seconds, which is the maximum length for most videos, indicating the participant watched the whole video and still got the wrong action perception. For fine videos, one expects this due to the great number of supervoxels being perceived,

16 *Chenliang Xu, Richard F. Doell, Stephen José Hanson, Catherine Hanson, and Jason J. Corso*
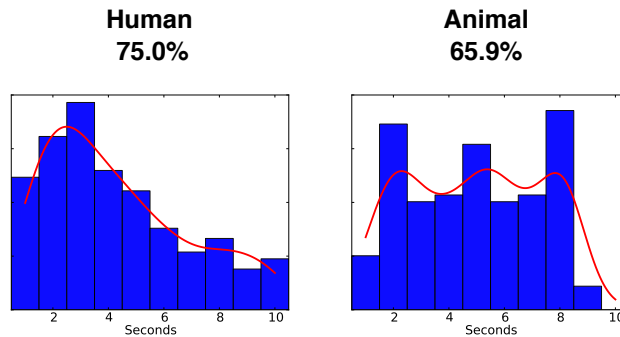

Fig. 6. Performance comparison between human actors and animal actors. The percentages on top are computed when the supervoxel human perceptions are correctly matched to ground truth ones. The response time plots include both correctly and incorrectly matched supervoxel human perceptions.

which introduces more noise. The medium and coarse scales are more uniformly distributed (although the coarse scale also has a peak at eight seconds), indicating that sometimes the perception was simply wrong. This may either be due to intrinsic limitation of the supervoxels to retain some action semantics or due to the ambiguities of the specific videos in the data set, which, although we did try to avoid, are present in some few cases. Further study on this point is needed to better understand the source of the error.

4.3. *How does the semantic retention vary with actor?*

We stratify the accuracy of the matches according to the actor performing the action. Figure 6 shows the overall performance by human actors and animal actors. In general, human perception of human actors has higher match than that of animal actors. For speed, the response time of human actors has only one peak at around three seconds, while that of animal actors has multiple peaks, which may be due to the greater variation in appearance of animals in the data set than of humans. Moreover, human activity is easier to perceive than animal as studied by Pinto and Shiffrar [45]. Considering the result in Table 1, the result in Figure 6 also suggests a correlation between knowing the actor and recognizing the action correctly.

4.4. *How does the semantic retention vary with static versus moving background?*

Figure 7 shows the overall performance of the supervoxel human perception match for static background and moving background. Supervoxel human perceptions have higher match and shorter response time in the case of static background, as expected (the dominant actor is more easily picked out by the participant). The relatively
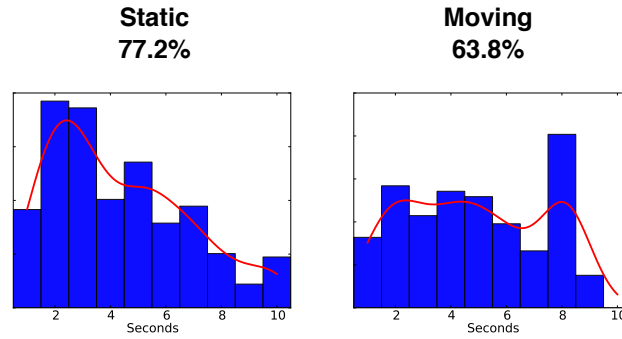
Fig. 7. Performance comparison between static background and moving background. The percentages on top are computed when the supervoxel human perceptions are correctly matched to ground truth ones. The response time plots include both correctly and incorrectly matched supervoxel human perceptions.
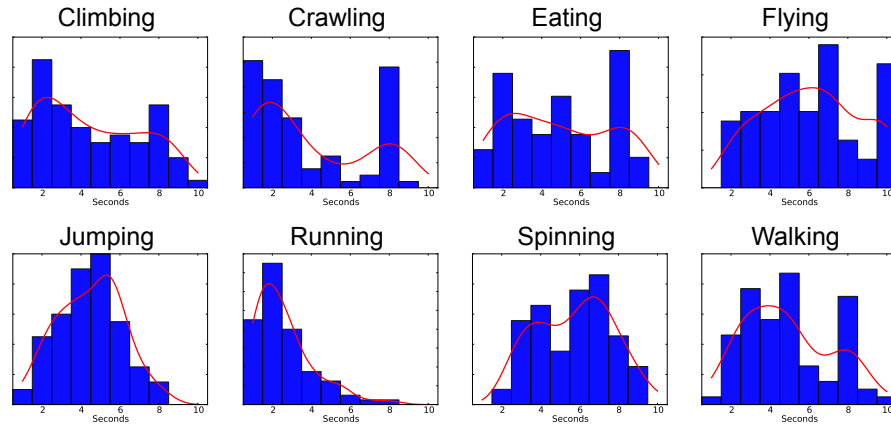

Fig. 8. Response time of eight different actions for both correctly and incorrectly matched supervoxel human perceptions.

"flat" curve in moving background suggests the response time for a single video highly depends on the specific background within that video.

4.5. *How does response time vary with action?*

Figure 8 shows the response time for the eight different actions. From the trend of the red curves in the figure, running and crawling get the shortest response time while the flying action takes longest. Bimodality in crawling is likely due to the very simple human baby crawling video (short response time) and very challenging cat preying video (long response time; see Figure 9 for the example). The more general
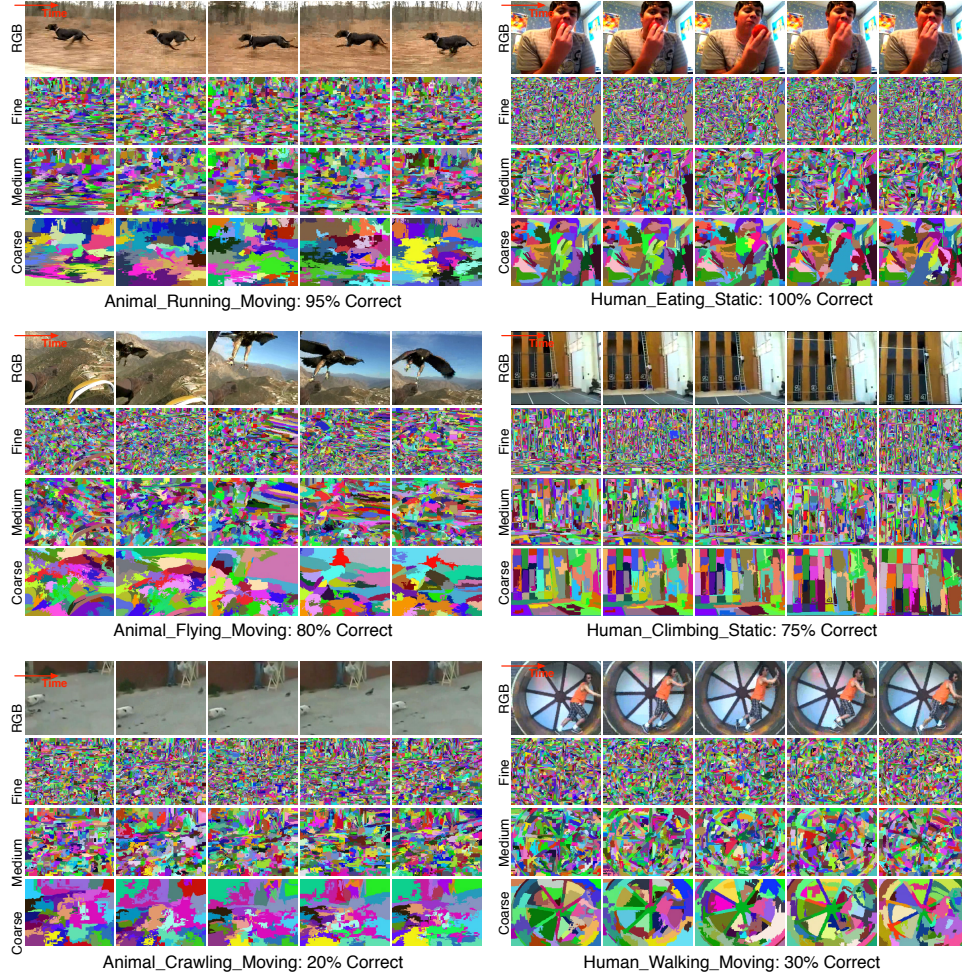
Fig. 9. Visualization of videos with different levels of semantic retention in human perception experiment. From top to bottom rows are videos picked from high, moderate, and low semantic retention resepectively. Frames are uniformly sampled from each video. We notice that supervoxel motion plays an important role in helping human observers locate the actor in a supervoxel segmentation video, which is hard to see in the montages. Therefore, we encourage people to view those videos in our project website for a better visualization.

messages behind these results are that those unusual actions such as human flying take more time to get a response, and that those actions whose semantics have been strongly retained (resulting in higher match statistics, Table 2) are generally responded to more quickly than those whose semantics have less well been retained.

A Study of Actor and Action Semantic Retention in Video Supervoxel Segmentation 19

	DT	HOG3D	Action Bank	SSC (Fine)	SSC (Median)	SSC (Coarse)
Actor	37.50%	37.50%	25.00%	51.65%	51.23%	**54.52%**
Action	18.75%	9.38%	**21.88%**	13.14%	10.11%	7.42%

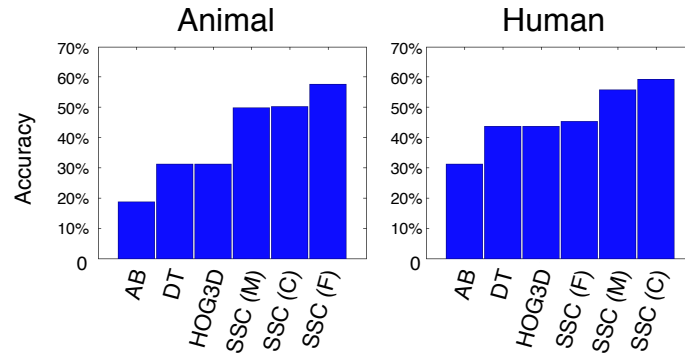Table 3. Overall machine classification accuracy. The leading score of each task is highlighted in bold font.

Fig. 10. Ranked classification accuracy on animal and human actors of different features.

4.6. *Easy, moderate and hard videos of human perception*

Finally, in Figure 9, we show montages of interesting videos, some with high action semantic retention and others with moderate or low retention. These top cases have distinct shape and motion properties that are readily transferred to the supervoxels; in the case of the running dog, the lateral motion is very strong. In the bottom left of retention examples, we see a cat crawling toward prey, but the cat is off-center from the camera and the participants likely dismiss this small off-center motion as noise for most of the video resulting in incorrect and slow responses. On the bottom right, we see a human walking in the spinning wheel. The human is walking; the wheel is spinning. There is likely semantic ambiguity here and further study is needed to understand the level and impact of the ambiguity.

4.7. *Machine Recognition Analysis*

Overall Performance. Table 3 shows the overall machine classification accuracy. Surprisingly, for actor classification, supervoxel shape context (SSC) gives much higher accuracy than the others. We suspect the reason why the performance of other features is worse than chance (50%) is because of a lack of discrimination in the feature (e.g., action bank has only human-based templates), or a lack of sufficient training samples in the data set. For action classificatoin, action bank achieves the leading performance, which is probably due to the underlying high-level semantics in the action templates. Note that the action classification by chance is 12.5%.

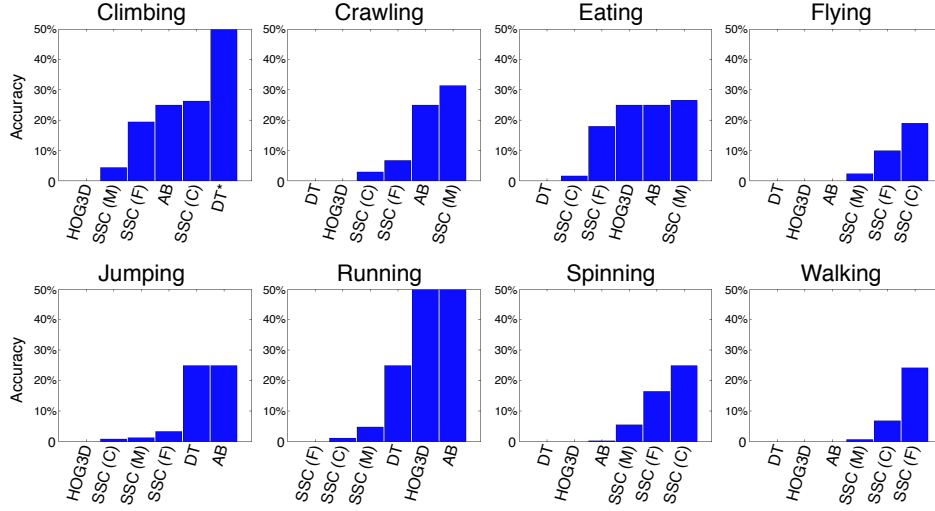**Actor.** Figure 10 shows the ranked classification performance on individual actors

Fig. 11. Ranked classification accuracy on eight actions separately. Note that DT* in the top-left plot achieves 100% accuracy on climbing.

(animal and human). Overall, the performance of human actor is higher than animal actor, which is similar to our findings of human perception in Section 4.1. For human actor, the performance increases when supervoxels in a video become coarser. We suspect that the shape information at coarser levels is cleaner than that at finer levels for training a human classifier. However, it dose not hold for animal actor, which we suspect is due to more irregular shapes of animals. Therefore more sophisicated methods are needed for finding the right scales for fine grained actors.

Action. Figure 11 shows the classification performance on eight actions. Different features have very different performance on actions. This is expected since the different types of feature have very distinct foci in a video. For example, action bank performs stably well for most of the actions, whereas DT has good performance only on climbing, jumping and running, which suggests that motion based feature is easier to capture unidirectional actions with limited training data. We also note that although DT has better classification accuracy in Table 3, supervoxel shape context achieves superior performance on more actions, such as crawling, eating, flying, spinning and walking. We show examples of success (spinning) and failure (running) cases of supervoxel shape context in Figure 12. We suspect the reason why supervoxel shape context fails to capture the running action is because the video only shows the upper body of the athlete.

In general, supervoxel shape context performs well on actor classification with a 54.52% accuracy at coarse level, but less well on action classification with a 13.14 % accuracy at fine level. The supervoxel shape information is essential for discriminate actors since animal and human are of very different shapes. Furthermore, more sophisticated methods are needed to better capture the shape changing information
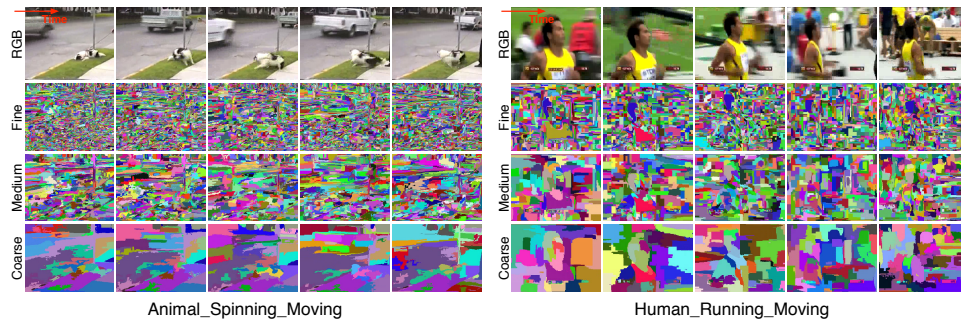
Fig. 12. Examples of success (left) and failure (right) cases of using supervoxel shape context to capture action.

which is crucial for action discrimination.

In summary, our experimental results demonstrate that the rich semantics in supervoxel segmentation can be well used in machine recognition to achieve competitive results when comparing with state of the art video level features. The overall machine recognition of actor and action seems to follow the same trend as that of human perception but more work needs to be done to get the machine recognition models up to par with the humans in terms of recognition performance.

5. Conclusion

In this paper, we explore the degree to which actor and action semantics are retained in video supervoxel segmentation. We design and conduct a systematic study to answer a set of questions related to this semantic retention. Our experiment results indicate strong retention of actor and action semantics: human perception achieves 82% accuracy on actor and 70% on action. Furthermore, this underlying semantics in supervoxel segmentation can be well used in machine recognition tasks. The supervoxel shape context achieves competitive results when comparing with state of the art video level features, but it is still far away from human performance. Our overall findings suggest that supervoxel segmentation is a rich decomposition of the video content, compressing the signal significantly while retaining enough semantic information to remain discriminative. The next challenge is to develop better methods for representation of the supervoxels for machine modeling.

Acknowledgements

References

[1] J. Tang, S. Yan, R. Hong, G.-J. Qi, and T.-S. Chua, "Inferring semantic concepts from community-contributed images and noisy tags," in *ACM International Conference on Multimedia*, 2009.

[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[3] A. Berg, J. Deng, S. Satheesh, H. Su, and L. Fei-Fei, "Imagenet large scale visual recognition challenge, http://www.image-net.org/challenges/LSVRC/2011/," 2011.

[4] J. Deng, A. Berg, K. Li, and L. Fei-Fei, "What does classifying more than 10,000 image categories tell us?" in *European Conference on Computer Vision*, 2010.

[5] T. Dean, M. A. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, and J. Yagnik, "Fast, accurate detection of 100,000 object class on a single machine," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[6] A. Tamrakar, S. Ali, Q. Yu, J. Liu, O. Javed, A. Divakaran, H. Cheng, and H. Sawhney, "Evaluation of low-level features and their combinations for complex event detection in open source videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[7] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, 2005.

[8] A. Klaser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *British Machine Vision Conference*, 2008.

[9] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[10] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[11] C. Xu, C. Xiong, and J. J. Corso, "Streaming hierarchical video segmentation," in *European Conference on Computer Vision*, 2012.

[12] O. Amir, I. Biederman, and K. J. Hayworth, "Sensitivity to nonaccidental properties across various shape dimensions," *Vision Research*, 2012.

[13] K. Grill-Spector, "The neural basis of object perception," *Current opinion in neurobiology*, 2003.

[14] J. F. Norman, F. Phillips, and H. E. Ross, "Information concentration along the boundary contours of naturally shaped solid objects," *Perception*, 2001.

[15] H. Öğmen, "A theory of moving form perception: Synergy between masking, perceptual grouping, and motion computation in retinotopic and non-retinotopic representations," *Advances in Cognitive Psychology*, 2007.

[16] C. Xu, R. F. Doell, S. J. Hanson, C. Hanson, and J. J. Corso, "Are actor and action semantics retained in video supervoxel segmentation?" in *IEEE International Conference on Semantic Computing*, 2013.

[17] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

[18] X. Ren and J. Malik, "Learning a classification model for segmentation," in *IEEE International Conference on Computer Vision*, 2003.

[19] S. Gould, R. Fulton, and D. Koller, "Decomposing a scene into geometric and semantically consistent regions," in *IEEE International Conference on Computer Vision*, 2009.

[20] W. Brendel and S. Todorovic, "Video object segmentation by tracking regions," in *IEEE International Conference on Computer Vision*, 2009.

[21] M. Grundmann, V. Kwatra, M. Han, and I. Essa, "Efficient hierarchical graph-based video segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[22] C. Xu and J. J. Corso, "Evaluation of super-voxel methods for early video processing," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[23] S. Palmer, *Vision science: photons to phenomenology.* MIT Press, 1999.

[24] M. Wertheimer, "Laws of organization in perceptual forms," *A source book of Gestalt psychology*, pp. 71–88, 1938.

[25] C. Xu, S. Whitt, and J. J. Corso, "Flattening supervoxel hierarchies by the uniform entropy slice," in *IEEE International Conference on Computer Vision*, 2013. [Online]. Available: http://www.cse.buffalo.edu/~jcorso/pubs/jcorso_ICCV2013_hieflat.pdf

[26] E. Sharon, M. Galun, D. Sharon, R. Basri, and A. Brandt, "Hierarchy and adaptivity in segmenting visual scenes," *Nature*, vol. 442, no. 7104, pp. 810–813, 2006.

[27] J. J. Corso, E. Sharon, S. Dube, S. El-Saden, U. Sinha, and A. Yuille, "Efficient multilevel brain tumor segmentation with integrated bayesian model classification," *Medical Imaging, IEEE Transactions on*, vol. 27, no. 5, pp. 629–640, 2008.

[28] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, 2004.

[29] S. Paris, "Edge-preserving smoothing and mean-shift segmentation of video streams," in *European Conference on Computer Vision*, 2008.

[30] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the nyström method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 214–225, 2004.

[31] F. Galasso, R. Cipolla, and B. Schiele, "Video segmentation with superpixels," in *Asian Conference on Computer Vision*, 2012.

[32] J. Chang, D. Wei, and J. W. F. III, "A video representation using temporal superpixels," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[33] M. Van den Bergh, G. Roig, X. Boix, S. Manen, and L. Van Gool, "Online video seeds for temporal window objectness," in *IEEE International Conference on Computer Vision*, 2013.

[34] M. Reso, J. Jachalsky, B. Rosenhahn, and J. Ostermann, "Temporally consistent superpixels," in *IEEE International Conference on Computer Vision*, 2013.

[35] G. Palou and P. Salembier, "Hierarchical video representation with trajectory binary partition tree," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[36] J. Tighe and S. Lazebnik, "Superparsing: scalable nonparametric image parsing with superpixels," *International Journal of Computer Vision*, 2010.

[37] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[38] S. H. Raza, M. Grundmann, and I. Essa, "Geometric context from videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[39] K. Tang, R. Sukthankar, J. Yagnik, and L. Fei-Fei, "Discriminative segment annotation in weakly labeled video," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[40] C. L. Zitnick and D. Parikh, "The role of image understanding in contour detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[41] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1986.

[42] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

24   *Chenliang Xu, Richard F. Doell, Stephen José Hanson, Catherine Hanson, and Jason J. Corso*

[43] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[44] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, pp. 1–11, 2012.

[45] J. Pinto and M. Shiffrar, "The visual perception of human and animal motion in point-light displays," *Social Neuroscience*, 2009.