# Adaptive Tuboid Shapes for Action Recognition

Roman Filipovych and Eraldo Ribeiro

Computer Vision and Bio-Inspired Computing Laboratory
Department of Computer Sciences
Florida Institute of Technology
Melbourne, FL 32901, USA
{rfilipov,eribeiro}@fit.edu

**Abstract.** Encoding local motion information using spatio-temporal features is a common approach in action recognition methods. These features are based on the information content inside subregions extracted at locations of interest in a video. In this paper, we propose a conceptually different approach to video feature extraction. We adopt an entropy-based saliency framework and develop a method for estimating tube-like salient regions of flexible shape (*i.e.,* tuboids). We suggest that the local shape of spatio-temporal subregions defined by changes in local information content can be used as a descriptor of the underlying motion. Our main goal in this paper is to introduce the concept of adaptive tuboid shapes as a local spatio-temporal descriptor. Our approach's original idea is to use changes in local spatio-temporal information content to drive the tuboid's shape deformation, and then use the tuboid's shape as a local motion descriptor. Finally, we conduct a set of action recognition experiments on video sequences. Despite the relatively lower classification performance when compared to state-of-the-art action-recognition methods, our results indicate a good potential for the adaptive tuboid descriptor as an additional cue for action recognition algorithms.

## 1 Introduction

Human action recognition has received significant attention in the computer vision community over the past decade. Action recognition is a challenging problem with a number of applications including surveillance, video-retrieval, and human-computer-interaction. In this paper, we address the issue of extracting descriptive features from motion videos. Inspired by object recognition methods [9,10,1], recent action recognition approaches have demonstrated the effectiveness of using local motion descriptors extracted at spatio-temporal locations across the video volume [4,11]. Local motion descriptors are usually built using filter responses or motion measurements calculated inside spatio-temporal subregions [7,4]. For example, Kläser *et al.* [7] calculate video descriptors from histograms of oriented gradients (HoG). Dollar *et al.* [4] consider spatio-temporal subregions of cuboid shape, and obtain descriptors using normalized pixel values, brightness gradients, and windowed optical flow.

G. Bebis et al. (Eds.): ISVC 2009, Part II, LNCS 5876, pp. 367–376, 2009.

Motion descriptors are usually extracted at the locations provided by spatio-temporal region detectors. For example, Laptev and Lidenberg [8] extended the Harris corner detector [5] to the spatio-temporal domain. In [8], interest points are detected by analyzing spatio-temporal filter responses over increasing scales, where the scale of the operator kernel determines the scale of the spatio-temporal subregion. Dollar *et al.* [4] proposed a spatio-temporal corner detector by modifying the temporal component of the operator kernel.

Another class of approaches work on the adaptation of the entropy-based salient region detector originally introduced by Kadir and Brady [6]. Their method works by considering changes in local information content over different scales. An extension of this detector to video domain was recently introduced by Oikonomopoulos *et al.* [11].

Current rigid spatio-temporal regions (*i.e.,* cuboids, ellipsoids, cylinders) [11,7,4] do not allow for the use of regions' shape as a cue for video analysis. As a result, these regions may not be able to capture nontrivial motion variations due to human's articulated motion. Additionally, methods based on cuboid- or elliptic-shaped subregions strongly rely on the availability of descriptive information content inside the analyzed subregions. In fact, changes in human appearance, illumination, and viewpoint may compromise the descriptiveness of the subregions' content. At the same time, the shape of traditional spatio-temporal subregions carries little information about the motion's local spatial properties.

In this paper, we propose new video features that are designed to "follow" the local spatio-temporal information flow. We adopt the region saliency framework [6,11] and develop a method for estimating tube-like salient regions (*i.e.,* tuboids) of adaptive shape. We argue that the shape of the local spatio-temporal information flow is important to describe local motion. We show how features that are invariant to scale, and partially invariant to viewpoint changes can be extracted from videos. Our main goal in this paper is to introduce the concept of adaptive tuboid shapes as a local spatio-temporal descriptor. Our main idea is to use changes in local spatio-temporal information content to drive the tuboid's shape deformation, and then use the tuboid's shape as a local motion descriptor. Finally, we conduct a set of experiments on real motion videos, and show that our new adaptive-shape descriptors can be effective for action recognition.

The remainder of this paper is organized as follows. In Section 2, we review the spatio-temporal subregion saliency measure from [11]. In Section 3, we introduce our spatio-temporal subregions of flexible tube-like shape, and describe the tuboid parameters-estimation procedure. In Section 4, we develop a set of descriptors based on the shapes of the extracted tuboids. In Section 5, we describe the action learning and recognition method used in our paper. Experimental results are reported in Section 6, with the paper concluding in Section 7.

## 2 Measuring Spatio-Temporal Information Content

In this section, we describe the saliency measure introduced by Kadir and Brady [6], and extended to the spatio-temporal domain by Oikonomopoulos *et al.* [11].

The method begins by calculating Shannon's entropy of local image attributes (*e.g.,* intensity, filter response) inside cylindrical spatio-temporal volumes over a range of scales. This entropy is given by:

$$H_D(\mathbf{s}) = -\int_{\mathbf{q}\in D} p_D(\mathbf{q},\mathbf{s}) \log_2 p_D(\mathbf{q},\mathbf{s}) d\mathbf{q}, \tag{1}$$

where $p_D(\mathbf{q},\mathbf{s})$ is the probability density function (pdf) of the signal in terms of scale $\mathbf{s}$, and descriptor $\mathbf{q}$ which takes on values from descriptors in the video volume $D$. Here, the pdf can be approximated by a pixel intensity histogram or by a kernel-based method such as Parzen windows. In our case, we follow [11], and use a histogram of the values obtained from the convolution of the image sequence with a Gaussian derivative filter.

The spatio-temporal subregion $D$ is assumed to be extracted at the origin of the coordinate system (*i.e.,* at the spatio-temporal location $\mathbf{x} = (0,0,0)^\mathsf{T}$). The scales $\mathbf{s} = (s_1,\ldots,s_n)$ represent the size parameters of the analyzed volumes (*e.g.,* spatio-temporal cylinder's radius and length). Once the local entropy values are at hand, a set of candidate scales is selected for which the entropy $H_D$ has local maxima, i.e.,

$$S = \left\{ \mathbf{s} : \frac{\partial H_D(\mathbf{s})}{\partial \mathbf{s}} = 0,\ \frac{\partial^2 H_D(\mathbf{s})}{\partial^2 \mathbf{s}} < 0 \right\}. \tag{2}$$

A saliency metric, $Y_D$, as a function of scales $\mathbf{s}$, can be defined as:

$$Y_D(\mathbf{s}) = H_D(\mathbf{s}) W_D(\mathbf{s}), \qquad \forall(\mathbf{s}) \in S, \tag{3}$$

where, for candidate scales in $S$, the entropy values are weighted by the following interscale unpredictability measure defined via the magnitude change of the pdf as a function of scale:

$$W_D(\mathbf{s}) = \sum_i s_i \int_{\mathbf{q}\in D} \left| \frac{\partial}{\partial s_i} p_D(\mathbf{q},\mathbf{s}) \right| d\mathbf{q}, \qquad \forall(\mathbf{s}) \in S. \tag{4}$$

An important property of the saliency measure in Equation 3 is that it does not depend on the content inside a subregion. Instead, the measure is based on changes in information content over scales. This makes the saliency metric $Y_D$ particularly robust. Next, we introduce our tuboid regions used in this paper, and propose a tuboid parameters estimation algorithm.

## 3 Tuboids

Spatio-temporal subregions considered in [11] have very simple shape (*i.e.,* cylinder). In this section, we propose to use subregions of more complex shape by introducing parametrization for tube-like video volumes.

Our tuboid model consists of a disk of variable radius that slides along a curve describing the temporal evolution of the spatial information content. We assume

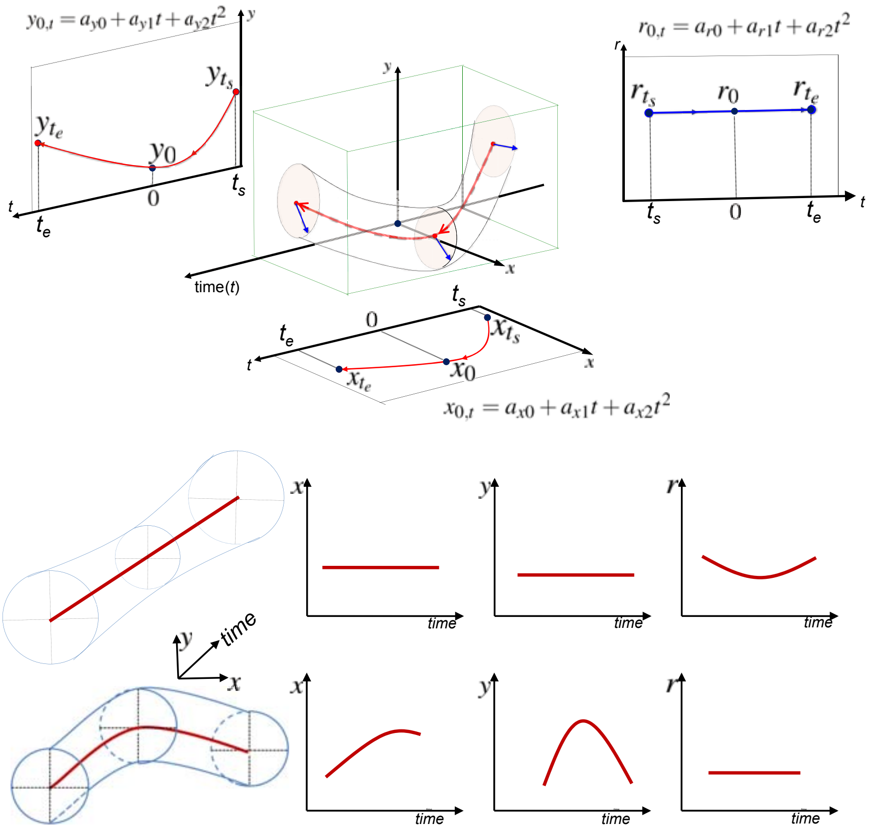

**Fig. 1.** Examples of tuboids and their parametric components as a function of time. Top: orthogonal projection curves describing the spatio-temporal variation of $(x, y)$ coordinates, as well as the temporal variation of the flexible cylinder radius. Bottom: examples of tuboids.

that a spatio-temporal subregion is centered at the origin of the coordinate system. At time $t$, a sliding disk $D_t$ of radius $r_t$ is given by:

$$(\mathbf{p} - \mathbf{c}_t) \cdot (\mathbf{p} - \mathbf{c}_t) \leq r_t^2, \tag{5}$$

where $\mathbf{p} = (x_t, y_t)^\mathsf{T}$ is a point on the disk, and $\mathbf{c}_t = (x_{0,t}, y_{0,t})^\mathsf{T}$ is the disk's center point. Let $\mathbf{g} = (x_{0,t}, y_{0,t}, r_t)^\mathsf{T}$ represent our tuboid model. We model the temporal evolution of tuboid $\mathbf{g}$ (*i.e.*, medial-axis points and disk radius) using quadratic parametric equations given by:

$$\mathbf{g}_t = \sum_{k=0}^{2} \mathbf{a}_k \, t^k \qquad \text{for} \qquad t \in [t_s, t_e]\,, \tag{6}$$

where $\mathbf{a}_k = (a_{xk}, a_{yk}, a_{rk})^\mathsf{T}$. The time values $t$ belong to a bounded interval starting at $t_s$ and ending at $t_e$. Equation 6 defines the tuboid's shape. Figure 1 shows examples of shapes described by Equations 5 and 6. In the case of cylindrical subregions considered in [11], the components in (6) are independent of time, and carry little information about the motion's local properties. Please, notice that the point $(x_{0,0}, y_{0,0})$ (*i.e.,* the medial-axis point for $t = 0$) may lie outside the tuboid, and thus will not necessarily coincide with the local coordinate system's origin. This characteristic is illustrated in Figure 1(top row). The allowed deviation of the axial curve from the local coordinate origin is controlled by the parameter estimation procedure, and will be discussed later in this paper. Examples of tuboids of different shapes are shown in the last two rows of Figure 1. The figure also shows plots of individual components of Equation 6.

Our goal is to estimate the parameters of a salient subregion described by (5) and (6). However, a direct optimization over the saliency measure $Y_D$ may not be applicable due to several reasons. First, the increase in subregion's dimensionality creates the possibility of degenerate cases [6]. Secondly, due to noise and space discretization, the set of candidate scales as described in (2) may be empty if the number of shape parameters is large. Finally, an exhaustive search over tuboid parameters is computationally intensive. We address these issues by using an alternative scale function, and by employing a gradient-ascent approach.

### 3.1 Estimating Tuboid Parameters

Following Equation 6, a tuboid can be completely defined by a set of eleven parameters, $\Theta = (a_{x0}, a_{x1}, a_{x2}, a_{y0}, a_{y1}, a_{y2}, a_{r0}, a_{r1}, a_{r2}, t_s, t_e)$. Unfortunately, obtaining a set of candidate scales for the set $S$ described in Equation 2 may not be achievable. Indeed, in our implementation we noticed that even for three parameters the set of candidate scales $S$ is often empty. To proceed, we define a scale function $F(\Theta)$ that is independent of the parameters of the tuboid's axial curve. This independence allows us to adapt the shape of the axial curve solely based on the tuboid's information content. The scale function is given by $F(\Theta) = r_{t_s} + r_0 + r_{t_e} + t_e - t_s$. Here, $F(\Theta)$ defines a tuboid's scale in terms of radii of disks obtained by $XY$-plane slices at times $t = t_s$, $t = 0$, and $t = t_e$, as well as by the tuboid's temporal length described by $t_s$ and $t_e$.

We begin with a set of initial tuboid parameters $\Theta_0$, and employ a gradient ascent search to maximize the saliency measure $Y_D$ over scales $F(\Theta)$. Assuming that parameters in $\Theta$ are discrete, the set of parameters at the next step of the gradient ascent algorithm is updated as:

$$\Theta_{n+1} = \Theta_n + \lambda \nabla Y_D \left( F(\Theta_n) \right) \tag{7}$$

where $\lambda$ is a small constant controlling the convergence speed of the search process. While in Equation 7 the direction of the search is determined by single-valued scales $F(\Theta_n)$, region $D$ during the iteration $n+1$ is defined by $\Theta_n$.

*Initialization.* In order to avoid local minima, the estimation procedure in Equation 7 requires good initialization of $\Theta_0$. We initialize the algorithm with the parameters of a cylindrical subregion that corresponds to the maximal change in information entropy as defined by Equation 1. The initialization procedure is given by the following maximization:

$$\Theta_0 = \arg\max_{\Theta} H_D\left(F(\Theta)\right),$$

$$\textbf{s.t.}, \quad \frac{\partial H_D(F(\Theta))}{\partial F(\Theta)} = 0, \quad \frac{\partial^2 H_D(F(\Theta))}{\partial^2 F(\Theta)} < 0, \quad x_t = 0, \quad y_t = 0, \quad r_{t_s} = r_0 = r_{t_e} \tag{8}$$

In (8), the conditions $x_t = 0$, $y_t = 0$, and $r_{t_s} = r_0 = r_{t_e}$, result in an initial spatio-temporal subregion of cylindrical shape. A crucial difference between our initialization procedure and the candidate scales estimation in Equation 2 is that we consider entropy changes over single-valued scales $F$. This is different from the joint maximization over individual parameters in (2). In this way, our parameters initialization approach is more likely to yield a solution. Nevertheless, for some spatio-temporal locations, the initialization in (8) may not result in a solution. In this case, the location is deemed non-salient and is removed from further consideration. The percentage of these points is small, due to the use of an interest detector pre-processing step that we will discuss later in this paper.

## 4 Feature Extraction

*General Features.* The quadratic curves described by Equation 6 can be characterized by three points. For simplicity, we choose the triplets $(x_{t_s}, x_0, x_{t_e})$ and $(y_{t_s}, y_0, y_{t_e})$ to define the axial curve's projections onto $XT$ and $YT$ planes, respectively. We also choose the triplet $(r_{t_s}, r_0, r_{t_e})$ to define $r_t$. We then define a general tuboid feature vector using eleven components as $\mathbf{d}_g = (x_{t_s}, x_0, x_{t_e}, y_{t_s}, y_0, y_{t_e}, r_{t_s}, r_0, r_{t_e}, t_s, t_e)$. While the general features $\mathbf{d}_g$ completely describe the shapes of the underlying tuboids, they may not be suitable for scenarios with varying camera parameters. Next, we propose a scale-invariant tuboid descriptor that is less sensitive to scale variations.

*Scale-Invariant Features.* We commence by assuming that the videos are obtained by static cameras with different scale parameters. The spatial components of scale-invariant tuboid shape descriptors can be obtained using unit vectors defined by the points $P_{t_s} = (x_{t_s}, y_{t_s})$, $P_0 = (x_0, y_0)$, and $P_{t_e} = (x_{t_e}, y_{t_e})$. The components of the scale-invariant descriptor are given by:

$$\mathbf{u} = \frac{\overline{P_{t_s} - P_0}}{\|\overline{P_{t_s} - P_0}\|}, \quad \mathbf{v} = \frac{\overline{P_0 - P_{t_e}}}{\|\overline{P_0 - P_{t_e}}\|}, \quad \text{and} \quad \mathbf{w} = \frac{\overline{P_{t_e} - P_{t_s}}}{\|\overline{P_{t_e} - P_{t_s}}\|}. \tag{9}$$

Examples of vectors $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ are shown in Figure 2. Additionally, scale effects on $r_t$ can be reduced by considering the additional feature components $\xi = r_{t_s}/r_0$ and $\zeta = r_{t_e}/r_0$. The scale-invariant tuboid descriptor is given by:

$$\mathbf{d}_s = (u_1, u_2, v_1, v_2, w_1, w_2, \xi, \zeta, t_s, t_e) \tag{10}$$

where $\mathbf{u} = (u_1, u_2)$, $\mathbf{v} = (v_1, v_2)$, and $\mathbf{w} = (w_1, w_2)$. The two temporal components $t_s$ and $t_e$ can be incorporated into $\mathbf{d}_s$ since scale variations do not affect the temporal evolution of the motion.

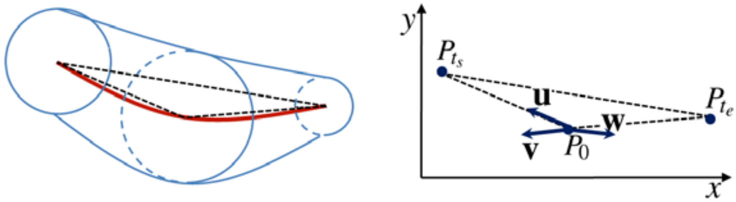

**Fig. 2.** Scale-invariant components $\mathbf{u} = (u_1, u_2)$, $\mathbf{v} = (v_1, v_2)$, and $\mathbf{w} = (w_1, w_2)$

*Viewpoint Invariant Features.* Here, we assume a simplified camera-viewing geometry in which camera parameters are limited to vertical or horizontal rotations about its center. We further assume that all cameras face the scene of interest, and are located at the same distance from it. Additionally, the scene is assumed to be far from the camera centers. These assumptions allow us to approximate projective camera transformations by affine transformations. In this paper, we consider descriptors that are invariant to horizontal or vertical circular translations of the camera (*i.e.,* $X$-view-invariant and $Y$-view-invariant features). The $X$-view-invariant feature vector consists of five components and is given by $\mathbf{d}_{xv} = (y_{t_s}, y_0, y_{t_e}, t_s, t_e)$. Similarly, it is possible to consider camera vertical circular translation. In this case, the $Y$-view-invariant feature vector is given by: $\mathbf{d}_{yv} = (x_{t_s}, x_0, x_{t_e}, t_s, t_e)$. The radial components $(r_{t_s}, r_0, r_{t_e})$ are not considered as they are not view-invariant.

## 5 Learning and Recognition

Our recogntion method is described as follows. Let $(\mathbf{d}^1, \ldots, \mathbf{d}^M)$ be a set of descriptors extracted from training video sequences of a specific action. Descriptors $\mathbf{d}^j$ are assumed to be of the same type (*e.g.,* general, scale-invariant, $X$-view-invariant, or $Y$-view-invariant). We model distribution of descriptors as a mixture of $K$ Gaussian densities given by $p(\mathbf{d}|\theta) = \sum_{i=1}^{K} \gamma_i \, p_i(\mathbf{d}|\theta_i)$, where $\mathbf{d}$ is a descriptor, $\theta_i$ are the parameters of $i$-th mixture, $\gamma_i$ represent the mixing weights such that $\sum_{i=1}^{K} \gamma_i = 1$, $p_i$ is a multivariate Gaussian density function parametrized by $\mu_i$ and $\Sigma_i$ (*i.e.,* the mean and covariance matrix, respectively), and $\theta$ presents the set of model parameters $(\gamma_1, \ldots, \gamma_K, \mu_1, \Sigma_1, \ldots, \mu_K, \Sigma_K)$. The parameters can be estimated by using the Expectation-Maximization (E.M.) algorithm [3]. Given a set of features $(\widehat{\mathbf{d}}^1, \ldots, \widehat{\mathbf{d}}^M)$ extracted from a novel video sequence, the classification score is given by $\tau = \frac{1}{M} \sum_{i=1}^{M} p(\widehat{\mathbf{d}}^i|\theta)$.

## 6 Experiments

The goal of our experiments is to demonstrate the potential of our tuboid-shaped descriptor. For this, we performed a set of classification experiments on the Weizmann human action dataset [2]. The dataset consists of videos of nine actions performed by nine individuals under similar camera conditions. Estimating tuboid parameters for every spatio-temporal location in a video is computationally intensive. Instead, we used an off-the-shelf spatio-temporal interest point detector [8,4] to obtain a set of candidate interest locations. In particular, we report our results using the detector from [4]. We applied a simple motion filter by convolving the image sequence with a derivative of a Gaussian along the temporal domain. The preprocessed sequences where then used to estimate tuboids' parameters at the locations of interest. In our experiments, we noticed that about 10% of the tuboids estimated at the provided locations had cylindrical shape (*i.e.,* $x_t$, $y_t$, and $r_t$ did not change over time). Cylindrical tuboids represent degenerate cases for scale-invariant features due to division by zero in Equation 9. As a result, we discarded those tuboids where the values $x_t$, $y_t$, and $r_t$ had small variances.

We first extracted general features $\mathbf{d}_g$ from the estimated tuboids, and adopted a leave-one-out scheme for evaluation by taking videos of actions performed by one individual for testing, and using sequences of the remaining individuals for training. The recognition rate achieved using these features was 85.2%. Similarly, we extracted scale-invariant features $\mathbf{d}_s$, $X$-view-invariant features $\mathbf{d}_{xv}$, and $Y$-view-invariant features $\mathbf{d}_{yv}$. Recognition rates were 75.3%, 61.7%, and 75.3% for scale-, $X$-view-, and $Y$-view-invariant features, respectively.

Next, we assessed the effect of the camera parameters changes on the performance of our descriptors. The main goal of this experiment is to show the robustness of the proposed approach in the presence of camera parameters variations. However, an interest point detector will generate different sets of interest locations under different camera conditions. As a result, using videos with uncontrolled scale (or viewpoint) variations does not allow to eliminate the influence of the interest region detector. To circumvent this issue, we resorted to a simulated change-of-scale approach. We rescaled frames in the original videos from the Weizmann dataset. For every video, a random scale was selected from the interval $[0.5, 2]$, and all frames in the sequence were rescaled to the same size. Additionally, the interest locations obtained for the original sequences were transformed accordingly. In this way, the effect of interest point detector was removed from our experiments. Examples of transformations considered in this paper are shown in Figure 4(c). Again, we extracted general, scale-invariant, $X$-view-invariant, and $Y$-view-invariant descriptors, and performed a leave-one-out validation for every descriptor type.

Finally, we simulated horizontal camera translations by rescaling the videos' horizontal dimension to a random scale. The performance of different features was assessed in the classification task. Similar experiment was performed for vertically rescaled sequences. The obtained recognition results are shown in Figure 4(d). The graph shows the recognition performance of our proposed

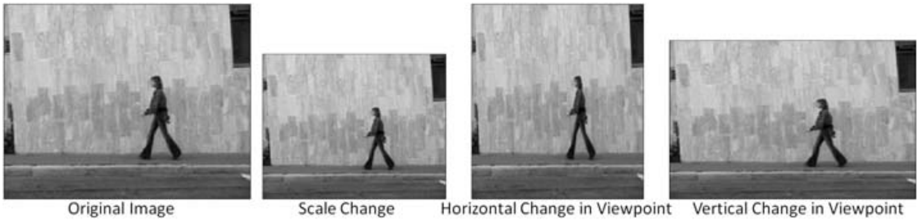

**Fig. 3.** Effects of transformations considered in our experiments

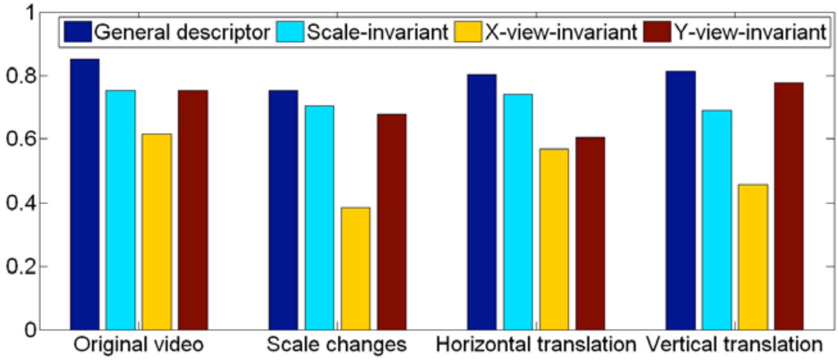

**Fig. 4.** Classification performance of our features under different camera conditions

features under different camera conditions. The plot suggests that tuboid general features allow for superior action recognition performance for all considered camera parameters. This effect is primarily due to the probabilistic nature of our learning method. Using $X$-view-invariant features resulted in the worst performance in our experiments. At the same time, $Y$-view-invariant features allow for a much better recognition performance as compared to $X$-view-invariant features. This suggests that, for the actions in the Weizmann dataset, the $X$-component of the motions is more descriptive than the $Y$-component (*i.e.,* viewpoint changes in the horizontal plane affect recognition performance more significantly than viewpoint changes in the vertical plane).

## 7 Conclusion

In this paper, we proposed a novel approach to video feature extraction. Rather than using information inside video-subregions, our features are based on the shapes of tuboid regions designed to follow the local information flow. We developed a set of general descriptors based on tuboid shapes, as well as scale-invariant, and partially view-invariant descriptors. Preliminary experiments performed on the Weizmann dataset suggest that the descriptor works well but there are a number of issues that still need attention. Among these issues is the need for a more comprehensive evaluation of the method on other motion datasets containing actual viewpoint and scale invariance. Additionally, our best recognition result was only 85.2% while state-of-the-art action recognition methods have achieved 100% recognition on the same dataset used in our experiments. This might be in part because our descriptor uses shape information only. While the tuboid's shape seems to be useful, we might be able to improve classification performance by combining the shape and content in the salient region. Finally, the temporal slices of our tuboids are of circular shape. By allowing a more flexible parametrization in Equation 5, it might be possible to include more information into the tuboid shape descriptors. We are currently working on these issues and the results of these studies will be reported in due course.

**Acknowledgments.** Research was supported by U.S. Office of Naval Research under contract: N00014-05-1-0764.

## References

1. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Comput. Vis. Image Underst. 110(3), 346–359 (2008)
2. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: ICCV, pp. 1395–1402 (2005)
3. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society Ser. B 39 (1977)
4. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS (October 2005)
5. Harris, C., Stephens, M.: A combined corner and edge detection. In: Proceedings of The Fourth Alvey Vision Conference, pp. 147–151 (1988)
6. Kadir, T., Brady, M.: Scale saliency: a novel approach to salient feature and scale selection. In: VIE, pp. 25–28 (2003)
7. Kläser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: British Machine Vision Conference, September 2008, pp. 995–1004 (2008)
8. Laptev, I., Lindeberg, T.: Space-time interest points. In: ICCV, Nice, France (October 2003)
9. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60(2), 91–110 (2004)
10. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. IJCV 60(1), 63–86 (2004)
11. Oikonomopoulos, A., Patras, I., Pantic, M.: Human action recognition with spatiotemporal salient points. IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics 36(3), 710–719 (2006)