

Human Action Recognition via Sum-Rule Fusion of Fuzzy K-Nearest Neighbor Classifiers

Teck Wee Chua Karianto Leman Nam Trung Pham

Institute for Infocomm Research

A*STAR (Agency for Science, Technology and Research), 1 Fusionopolis Way, Singapore
{tewchua, karianto, ntpham}@i2r.a-star.edu.sg

Abstract—Shape and motion are two most distinct cues observed from human actions. Traditionally, K-Nearest Neighbor (K-NN) classifier is used to compute crisp votes from multiple cues separately. The votes are then combined using linear weighting scheme. Usually, the weights are determined in a brute-force or trial-and-error manner. In this study, we propose a new classification framework based on sum-rule fusion of fuzzy K-NN classifiers. Fuzzy K-NN classifier is capable of producing soft votes, also known as fuzzy membership values. Based on Bayes theorem, we show that the fuzzy membership values produced by the classifiers can be combined using sum-rule. In our experiment, the proposed framework consistently outperforms the conventional counterpart (K-NN with majority voting) for both Weizmann and KTH datasets. The improvement may attribute to the ability of the proposed framework to handle data ambiguity due to similar poses present in different action classes. We also show that the performance of our method compares favorably with the state-of-the-arts.

Index Terms—Action recognition, Sum-Rule Fusion, Fuzzy K-NN

I. INTRODUCTION

Human can recognize an action in a seemingly effortless fashion. In contrast, the solutions using computer vision have, in many cases, proved to be immensely difficult. The challenges include: huge collection of possible actions, poor recording settings, temporal variations, inter/intra-personal differences etc. In addition, the choice of optimal representation for an action is still an open problem. Recent approaches can be categorized into global and local representations. For global representation, the region-of-interest (obtained by tracking or background subtraction) is encoded as a whole. In other words, the entire human figure is considered. Silhouette, optical flow, edge and space-time volume fall into this category. Efros *et al.* [1] used blurred optical flows to recognize the actions of small human figures. Blank *et al.* [2] stacked silhouettes over a sequence to form space-time volumes. Poisson equation was used to compute local space-time saliency and orientation features. Ikizler *et al.* [3] extended the motion descriptor of Efros by using spatial and directional binning and then combined it with line shape descriptor. Following that, they proposed to use histogram of oriented rectangles as the shape descriptor [4]. Likewise, Lin *et al.* [5] used silhouettes as the shape descriptor by counting the number of foreground pixels and motion-compensated optical flow as motion descriptor. On the contrary, local representation encodes the image or video

frames as a collection of local patches. Common local representation includes spatio-temporal interest points such as 3D Harris, cuboid, Hessian, Dense etc. Usually the interest points are extracted at different spatial and temporal scales. Laptev and Lindeberg [6] proposed to extend Harris corner detector to the third dimension. Dollár *et al.* [7] cuboid detector is based on temporal Gabor filter. Willems *et al.* [8] measured the saliency with the determinant of the 3D Hessian matrix.

According to Ikizler *et al.* [4], an action is defined by three key elements: (a) pose of the body, (b) speed of body motion, and (c) relative ordering of the poses. Actions such as jogging, walking and running can be easily confused if only pose information is used due to similar postures in the action sequences. It has been suggested that combinations of multiple cues may overcome the limitations of a single representation [9]. This is because different inputs may offer complimentary information about the patterns to be classified. In view of this, previous studies which used multiple cues were mainly based on linear weighting of classifiers' votes [3], feature vectors concatenation [5], [10], and hierarchical system [4], [3], [11]. While linear weighting scheme may seem straightforward, the weights are usually obtained in a brute-force or trial-and-error manner.

Apart from the uncertainty in determining the correct weight, most of the previous studies relied on conventional K-nearest neighbor classifier [12]. However, the algorithm has a major drawback: each prototype is considered equally important in the assignment of input patterns. Atypical data are weighted as much as the data from the true class. This can cause ambiguity in those places where the data overlap.

In this work, we address the aforementioned problems in two ways. First, we propose to use fuzzy K-NN classifiers to classify shape and motion feature respectively. The outputs of the classifiers can be regarded as the confidence levels of the input belonging to different classes. Next, we show that based on Bayes theorem the outputs of the classifiers can be fused using sum-rule.

The paper is organized as follows. In Section II we give an overview of the proposed action recognition framework. Next, we describe the details of shape and motion feature extraction in Section III. Subsequently, we outline the fuzzy K-NN algorithm and propose the sum-rule as the fusion strategy of the fuzzy outputs in Section IV. Section V sets the backdrop

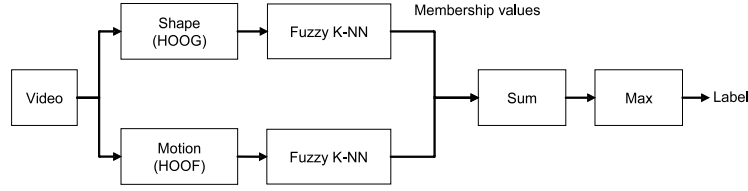


Fig. 1. The proposed shape-motion path combination strategy using Fuzzy K-NN with sum-rule.

for the experimental evaluation while Section VI shows the results with some comparisons to the state-of-the-art methods. Finally, Section VII gives the conclusion remarks of the paper.

II. SYSTEM OVERVIEW

The proposed system is illustrated in Fig. 1. From the shape path, fuzzy K-NN classifier produces membership value $\mu_i = \{\mu_{i1}, \mu_{i2}, \dots, \mu_{iC}\}$ for each snippet where C denotes the number of classes. Given a video sequence with M snippets, we then compute the average of the membership values from all snippets, i.e., $U_s = (\sum_i \mu_i)/M = \{U_1, U_2, \dots, U_C\}$ and normalize it to $\tilde{U}_s \in [0, 1]$. Likewise, we compute the normalized membership values $\tilde{U}_m$ from the motion path. Next, we combined the membership values from both shape and motion inputs using sum-rule

$$U = \tilde{U}_s + \tilde{U}_m \quad (1)$$

Finally, we apply winner-takes-all principle to determine the class label.

Since the proposed system will be compared with the conventional system (K-NN with majority voting), we briefly discuss the architecture of the latter (see Fig. 2). Consider the shape path first, for each test snippet we find the K-nearest snippets from the training set and assign its label to the class most common among its neighbors. We can obtain crisp votes $v_i = \{v_{i1}, v_{i2}, \dots, v_{iC}\}$ from the i -th snippet. The votes from the individual snippets (from the same video) are then summed up, i.e., $V_s = \sum_i v_i = \{V_1, V_2, \dots, V_C\}$. The step is repeated for the motion path to obtain the votes V_m . The crisp votes from the shape and motion paths are combined using linear weighting scheme

$$V = \alpha V_s + (1 - \alpha) V_m \quad (2)$$

where $\alpha \in [0, 1]$ is a design parameter that adjust the relative importance of shape and motion votes. Finally, the video is assigned to the class that has the highest votes, V .

III. MOTION AND SHAPE FEATURE EXTRACTION

As delineated in the introduction part, observing shape and motion is a very natural way to recognize an action. Motivated by the robustness of the histogram of feature, we use histogram-of-oriented gradient (HOOG) and histogram-of-oriented optical flow (HOOF) as the shape and motion descriptors respectively. We adopt the histogram formation method which was originally introduced by Chaudhry *et al.* [13]. The method is more robust against scale variation and the change of motion direction. The method is illustrated in Fig. 3(a) with

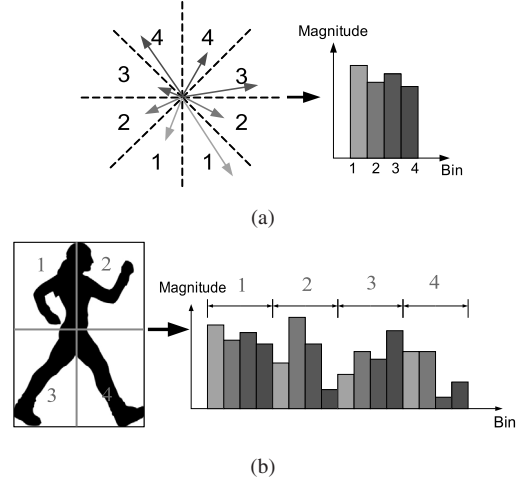


Fig. 3. (a) Histogram binning, (b) the bounding box is divided into 4×4 grid and the resultant histograms from each region are concatenated.

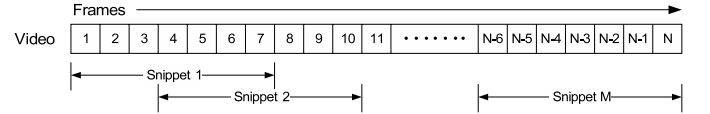


Fig. 4. An example of video sequence that is divided into overlapping snippets.

example of creating a 4-bin histogram. The main idea is to bin the vectors according to their primary angles from the horizontal axis. Therefore, the vectors are symmetry about the vertical axis. As a result, the histogram of a person moving from left to right will be same as the one with a person moving in the opposite direction. The contribution of each vector is proportional to its magnitude. The histogram is normalized to sum up to unity to make it scale-invariant. Therefore, we do not normalize the size of the bounding box. We further enhance Chaudhry *et al.* algorithm by including spatial information. This is done by dividing the bounding box of the subject into 4×4 regions as shown in Fig. 3(b).

The pose ordering and time-warping effect can be handled by using overlapping snippets instead of a single frame. Fig. 4 shows an example of 7-frame snippet with four overlapped frames, which is the setting that we used for all experiments in this study.

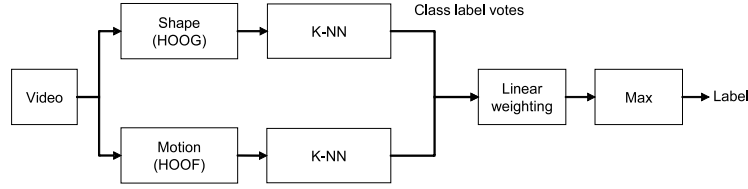


Fig. 2. The conventional shape-motion path combination strategy using K-NN with linear weighting.

IV. SUM-RULE FUSION OF FUZZY K-NEAREST NEIGHBOR CLASSIFIERS

Unlike the conventional crisp K-NN, fuzzy K-NN algorithm [14] assigns class membership value to an input data rather than class label. The membership assignment is based on the distance between the input data and its K-nearest neighbors and those neighbors' memberships (pre-computed during initialization process) in the possible classes. Let $S = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ be a set of n labeled samples. Also, let $\mu_j(\mathbf{x})$ be the membership assigned to the input data $\mathbf{x}$, and $\mu_{ij}(\mathbf{x})$ be the membership in the j th class of the i th labeled prototype.

$$\mu_j(\mathbf{x}) = \frac{\sum_{i=1}^K \mu_{ij} \left(1/\hbar(\mathbf{x}, \mathbf{x}_i)^{2/(m-1)} \right)}{\sum_{i=1}^K \left(1/\hbar(\mathbf{x}, \mathbf{x}_i)^{2/(m-1)} \right)} \quad (3)$$

Note that in this work $\hbar(\mathbf{x}, \mathbf{x}_i)$ denotes Hellinger distance instead of Euclidean distance [14] because the former provides a more accurate histogram-type feature similarity measure. The parameter m ($m > 1$) is used to scale the effect of the distance measure. The algorithm requires the prototype membership value to be initialized:

$$\mu_{ij}(x) = \begin{cases} 0.51 + (n_j/K) * 0.49 & \text{if } j = i \\ (n_j/K) * 0.49 & \text{if } j \neq i \end{cases} \quad (4)$$

where n_j is the number of the neighbors which belong to the j th class. This initialization procedure ensures the prototype is more informative. The more accurate the initial membership value is, the more accurate the classification result will be.

Suppose there are R classifiers and each r th classifier will classify a distinct input vector $\mathbf{x}_r$, ($r = 1, \dots, R$). Bayes rule can be applied to find the class label, c_{j^*} :

$$j^* = \arg \max_j (P(c_j | \mathbf{x}_1, \dots, \mathbf{x}_R)) \quad (5)$$

where

$$P(c_j | \mathbf{x}_1, \dots, \mathbf{x}_R) = \frac{p(\mathbf{x}_1, \dots, \mathbf{x}_R | c_j) P(c_j)}{p(\mathbf{x}_1, \dots, \mathbf{x}_R)} \quad (6)$$

In order to evaluate (5), the joint probability density function $p(\mathbf{x}_1, \dots, \mathbf{x}_R | c_j)$, which is difficult to infer, must be computed. Based on two assumptions, sum-rule can be used to combine the outputs of multiple classifiers [9]. The assumptions are: (a) the inputs $\mathbf{x}_r$ are conditionally statistically independent, and (b) a posteriori probabilities of individual classifier $P(c_j | \mathbf{x}_r)$ will not deviate from the prior probabilities $P(c_j)$ by a large margin. The first assumption is readily satisfied because we are using two types of representations: shape and motion.

Moreover, the second assumption can be satisfied when the discriminatory information is highly ambiguous due to the high level of noise. For example, KTH dataset has very strong intra-subject variations due to different recording scenarios. The sum-rule is given as:

$$j^* = \arg \max_j \left[(1 - R)P(c_j) + \sum_{r=1}^R P(c_j | \mathbf{x}_r) \right] \quad (7)$$

When priors are equal, (7) can be reduced to:

$$j^* = \arg \max_j \left(\sum_{r=1}^R P(c_j | \mathbf{x}_r) \right) \quad (8)$$

For fuzzy K-NN, we can approximate $P(c_j | \mathbf{x})$ as:

$$P(c_j | \mathbf{x}) \approx \mu_j(\mathbf{x}) \quad (9)$$

Notice that unlike the estimation of posterior probability from Neural Network where normalization of output is required [9], there is no need to normalize the output of fuzzy K-NN classifier because $\mu_j(\mathbf{x}) \in [0, 1]$.

There are some other forms of classifiers combination strategies such as product-rule, min-rule, max-rule, and majority voting proposed by Kittler *et al.* [9]. We will compare the performance of sum-rule to these variants in Section VI.

V. EXPERIMENTS

We performed various experiments to evaluate the proposed action recognition framework on two publicly available datasets (see Fig. 5):

- **Weizmann** The dataset was originally introduced in [2]. The dataset contains 90 low-resolution (180×144 pixels) video sequences with 9 subjects performing 10 actions: bend (*bend*), jumping-jack (*jack*), jump-forward (*jump*), jump-in-place (*pjump*), run (*run*), gallop-sideways (*side*), jump-forward-one-leg (*skip*), walk (*walk*), wave-one-hand (*wave1*), and wave-two-hands (*wave2*)¹. We used the silhouettes provided to compute the bounding boxes for the subjects. HOOG and HOOF features are extracted from the silhouettes.
- **KTH** The dataset was introduced in [15]. There are 25 subjects performing 6 actions: boxing, handclapping, handwaving, jogging, running, and walking. The low resolution (160×120) videos were recorded under four

¹Note that there are two versions of Weizmann dataset, the original one has 9 actions while the augmented version has 10 which includes *skip* action.



Fig. 5. Examples of different actions from databases Weizmann (left) and KTH (right).

scenarios (s1- outdoors, s2- outdoors with scale variation, s3- outdoor with different clothes, s4- indoors with lighting variation) and each video was split into 4 sub-clips. Originally, the dataset has $(4 \text{ settings}) \times (25 \text{ subjects}) \times (6 \text{ actions}) \times (4 \text{ sub-clips}) = 2400$ clips. However, only 2391 clips are available because 8 clips were missing. We used the bounding box provided by Lin *et al.* [5] to locate the subject. Nevertheless, we did not compute the silhouette because object segmentation is not our focus in this work. Therefore, HOG and HOF features are extracted directly from the raw grayscale video frames.

In the literature, KTH dataset has been regarded either as one large set with strong intra-subject variations (*all-in-one*) or as four independent scenarios. In the latter case, each scenario is trained and tested separately. In this work, we only focus on the *all-in-one* case. Since KTH dataset size is much larger than Weizmann dataset, we use K-means algorithm to cluster the training data. Each class is quantized into 500 clusters. This can reduce the intra-class variation and computational time. Leave-one-out cross validation (LOOCV) protocol is used in all evaluations.

For Fuzzy K-NN, we set parameter $m = 2$ such that the influence of each neighbor is weighted by the square of reciprocal of its distance from the query point. The parameter K is varied from 3, then 5 to 7 to evaluate the effect of the neighborhood size.

VI. RESULTS AND DISCUSSIONS

From the experiment on Weizmann dataset using the proposed framework and the conventional framework, it is evident that neighborhood size of 3 consistently gives better performance than other sizes. Table I shows the LOOCV recognition rate for Weizmann dataset when using neighborhood of $K = 3$. By comparing the results using single representation (either shape or motion feature only), it is noticed that fuzzy K-NN always performed better than K-NN counterparts with improvement of 3.33% and 2.22% when using the shape and motion features respectively. When the outputs from fuzzy K-NN classifiers were combined using sum-rule, the proposed system yielded 100% accuracy. In contrast, the conventional K-NN with linear weighting scheme only achieved 97.78%. Note that for this conventional scheme, the weight, α was set to 0.5 so that the shape and motion features were equally important to facilitate a common comparison setting for both systems.

The result for KTH dataset is tabulated in Table II. It is worth noting that fuzzy K-NN consistently outperforms K-NN counterpart when using single representation only. The advantage of the proposed framework is even more promi-

TABLE I
WEIZMANN AVERAGE ACCURACY ($K=3$, LOOCV).

Framework	Shape	Motion	Combined
KNN + Linear weighting	92.22%	92.22%	97.78%
Fuzzy K-NN + Sum Rule	95.55%	94.44%	100%

nent when both shape and motion inputs are combined. The classification accuracy is improved by as much as 5% when $K = 3$. We also investigate the effect of neighborhood size K . The results show that $K = 3$ gives the best results for both frameworks. Considering all three neighborhood sizes, the proposed framework outperforms the conventional counterpart by 3.59% in classification accuracy on average.

TABLE II
KTH AVERAGE ACCURACY (*all-in-one*, LOOCV).

Framework	K	Shape	Motion	Combined
KNN + Linear weighting	3	74.22%	82.44%	85.80%
	5	72.73%	82.14%	85.46%
	7	72.57%	82.14%	85.12%
Fuzzy K-NN + Sum Rule	3	77.97%	83.13%	90.80%
	5	76.91%	82.25%	88.43%
	7	74.76%	83.59%	87.93%

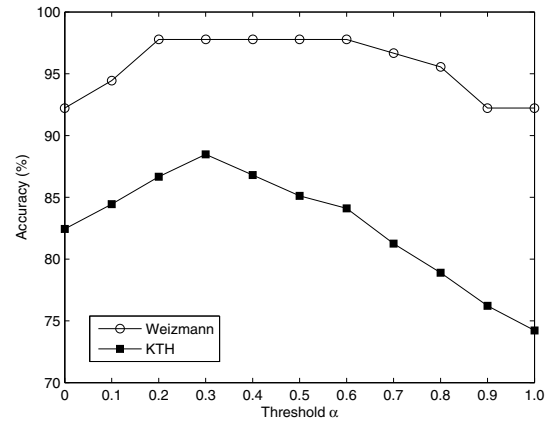


Fig. 6. The effect of weight α in the conventional linear weighting scheme on the average classification rate for Weizmann and KTH datasets. The weight is used to control the relative importance of the votes from shape and motion cues. $\alpha = 0$ corresponds to motion only while $\alpha = 1$ corresponds to shape only. The neighbor size, K is set to 3.

The results from Table I and II suggest that the contribution of each feature is dataset dependant. For instance, the shape feature is slightly better than the motion feature for



Fig. 7. Similar poses observed from *jogging* (left) and *running* (right) action sequences.

Weizmann dataset. On the other hand, we observe that for KTH dataset motion feature is much more discriminative than the shape feature. This implies that in the conventional linear weighting approach the contribution of each feature must be pre-determined by finding a suitable weight, α . Therefore, we evaluated the effect of α as shown in Fig. 6. The best results are only 97.78% for Weizmann and 88.48% for KTH. This provides clear evidence that regardless of the weight values, the conventional framework is still inferior to our implicitly weight-free framework. Another interesting observation from Table II is that the performance improvement brought by fuzzy K-NN classifier is more significant on the classification based on shape rather than motion features. The reason could be that fuzzy K-NN is able to handle the uncertainty in the poses better compared to conventional K-NN classifier. The uncertainty can arise from similar postures in action sequences such as jogging and running as shown in Fig. 7. Recall that fuzzy K-NN has an initialization procedure to weight the training sample which potentially resolves the data overlap problem. On the contrary, conventional K-NN suffers from the drawback that each training sample is considered equally important which causes atypical sample is weighted as much as the sample from the true class.

Finally, we investigate the performance of sum-rule compared to other combining schemes. The result from Table III clearly show that sum-rule delivers the best classification rate, closely followed by product-rule, max-rule, min-rule, and majority voting schemes. The result is consistent with Kittler *et al.* [9] findings that sum-rule produces the most reliable decisions due to its robustness against estimation error.

TABLE III
CLASSIFICATION RATE FOR KTH DATASET USING DIFFERENT
COMBINING SCHEMES

Combining Schemes	Accuracy (%)
Sum-rule	90.80
Product-rule	89.61
Max-rule	87.94
Min-rule	86.09
Majority vote	81.94

The confusion matrix for KTH is given in Fig. 8. Most of the errors occur when *running* action is misclassified as *jogging* action. This is followed by *boxing* and *clapping* actions. This observation is consistent with the results reported in [10], [8], [16]. Although the main goal of this work is to

	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing	87.88	5.05	3.03			4.04
Handclapping	2.02	91.92	5.05			1.01
Handwaving		4.00	95.00			1.00
Jogging				99.00	1.00	
Running				24.00	71.00	5.00
Walking						100.00

Fig. 8. Confusion matrix for KTH ($K = 3$).

TABLE IV
COMPARISON OF RECOGNITION RATES FOR WEIZMANN.

Method	Accuracy (%)
Our method	100.00
Fathi [17]	100.00
Blank [2]	99.64
Jhuang [11]	98.80
Wang [18]	97.78
Saad [19]	95.75
Chaudhry [13]	94.44
Niebles [20]	90.00

compare our framework to the conventional K-NN with linear weighting method, we also compare the results against state-of-the-art action recognition approaches. Table IV shows that our method achieved the same perfect accuracy as [17] for Weizmann dataset. As for KTH dataset, it may be argued that the results are not directly comparable as different authors employed different evaluation protocol (splits vs. LOOCV). While not definitive, Table V still provides some indicative comparison. The result shows that our system remains competitive compared to the other methods. In fact, our method achieved the second best result (only slightly inferior to Lin *et al.* [5]). One possible reason is that Lin *et al.* used silhouette for feature extraction. It is well known that silhouette requires good background modelling which is more restrictive than the bounding box-based approach. In contrast, our KTH experiment only used the original grayscale image containing inside the bounding box. From the result, it can be deduced that our method can perform very well even without using silhouette. In addition, it should be noted that in Lin *et al.* work the prototype tree is only used for quantization when matching the frame to prototype. The algorithm still relies on dynamic time warping which are computationally expensive when the video is long. Our framework has the advantage of efficient computation.

TABLE V
COMPARISON OF RECOGNITION RATES FOR KTH.

Method	Protocol	Accuracy (%)
Our method	LOOCV	90.80
Lin [5]	LOOCV	93.43
Fathi [17]	Splits	90.50
Ahmad [21]	Splits	88.83
Willems [8]	Splits	84.26
Niebles [20]	LOOCV	83.33
Dollár [7]	LOOCV	81.17
Ke [16]	LOOCV	80.90
Schüldt [15]	Splits	71.72

VII. CONCLUSION

This paper presents a new fuzzy-KNN with sum-rule fusion framework to perform action recognition from shape-motion features. The novelty is in the new approach to fuse the decisions from shape-motion cues which is very efficient in implementation. We demonstrated that fuzzy K-NN classifier can deliver better performance than the conventional K-NN classifier especially when dealing with data uncertainty. This is evident when classifying actions with similar poses. The introduction of sum-rule to combine fuzzy K-NN classifiers outputs has shown promising results compared to other combining schemes such as product-rule, max-rule etc. One major advantage of using sum-rule is that the process of choosing the suitable weight through trial-and-error is no longer required. The method described in this paper facilitates a generic framework that can also be used to combine the classification results from heterogenous inputs other than shape and motion.

REFERENCES

- [1] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. of International Conference on Computer Vision (ICCV)*, Nice, France, 2003.
- [2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. of International Conference on Computer Vision (ICCV)*, Beijing, China, 2005.
- [3] N. Ikizler, R. G. Cinbis, and P. Duygulu, "Human action recognition with line and flow histograms," in *Proc. of International Conference on Pattern Recognition (ICPR)*, Tampa, Florida, USA, 2008.
- [4] N. Ikizler and P. Duygulu, "Histogram of oriented rectangles: A new pose descriptor for human action recognition," *Image and Vision Computing*, vol. 27, no. 10, pp. 1515–1526, 2009.
- [5] Z. Lin, Z. Jiang, and L. S. Davis, "Recognizing actions by shape-motion prototype trees," in *Proc. of International Conference on Computer Vision (ICCV)*, Kyoto, Japan, 2009.
- [6] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proc. of International Conference on Computer Vision (ICCV)*, Nice, France, 2003.
- [7] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. of IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, Beijing, China, 2005.
- [8] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proc. of European Conference on Computer Vision (ECCV)*, Marseille, France, 2008.
- [9] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, 1998.

- [10] K. Schindler and L. Van Gool, "Action snippets: How many frames does human action recognition require?" in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, Alaska, 2008.
- [11] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," in *Proc. of International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brazil, 2007.
- [12] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, June 2010.
- [13] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, Florida, 2009.
- [14] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy K-NN neighbor algorithm," *IEEE Trans. Syst., Man, Cybern.*, vol. 15, no. 4, pp. 580–585, 1985.
- [15] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. of International Conference on Pattern Recognition (ICPR)*, Cambridge, UK, 2004.
- [16] Y. Ke, R. Sukthankar, and M. Hebert, "Spatio-temporal shape and flow correlation for action recognition," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Minnesota, USA, 2007.
- [17] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, Alaska, 2008.
- [18] L. Wang and D. Suter, "Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Minnesota, USA, 2007.
- [19] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 288–303, 2010.
- [20] J. C. Niebles, H. Wang, and L. Fei-fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int'l J. Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.
- [21] M. Ahmad and S. Lee, "Human action recognition using shape and CLG-motion flow from multi-view image sequences," *Pattern Recognition*, vol. 41, no. 7, pp. 2237–2252, July 2008.