HUMAN ACTION RECOGNITION USING HISTOGRAPHIC METHODS AND HIDDEN MARKOV MODELS FOR VISUAL MARTIAL ARTS APPLICATIONS

Sotirios Stasinopoulos and Petros Maragos

School of ECE, National Technical University of Athens, 15773 Athens, Greece

sotstas@gmail.com, maragos@cs.ntua.gr

ABSTRACT

Human Action Recognition is being used with an increasing rate in applications designed to describe human activity in everyday life. However, some areas still remain far from the epicenter of scientific research, like dynamic problems of detection and classification of movements from visual Martial Arts. With this paper, we are proposing a novel recognition system focused on these types of action, based on the use of local spatio-temporal features from Histographic methods, as those extracted by Histograms of Oriented Gradient (HOG) and Histograms of Optical Flow (HOF), while we also pursue the reduction of the problem's dimensionality by applying on them Principal Components Analysis (PCA). In continuation, we combine these features with the use of Hidden Markov Models (HMM) in order to train models for each different movement. Our system is tested with very encouraging results upon a database comprising sequences of shotokan karate movements (katas), created by us for the needs of this research. In parallel, we additionally attach an educational character to our application, with the extraction of a score for the accuracy of execution of each movement based on the prototypes we have built.

Index Terms— human action recognition, martial arts, karate, histogram of oriented gradient, histogram of optic flow, principal component analysis, hidden markov models

1. INTRODUCTION

With the rapid progress in computer vision and pattern recognition, a variety of video processing systems have been proposed to tackle a variety of visual problems, recognizing human activity in everyday applications, spanning from detection of human movement by surveillance cameras to recognition of sign language. Nevertheless, numerous areas of applications have not been adequately approached by the research community, regarding the type of recognition systems more suitable for them, their parameters and the further interaction with the user that could stem out of their analysis. Among them, actions belonging to Martial Arts, a complex and multi-variable category, provide an excellent paradigm of the extent to which Action Recognition can be implemented.

This paper proposes a novel human action recognition system particularly focused on detecting and classifying such complex dynamic actions, that may involve a large number of consequent stages, thus requiring the use of memory between them. The parameters of the system have been adjusted for application on various styles of Martial Arts, during which an athlete is required to execute a strictly defined and consistent sequence of movements, with great attention to the proper orientation of the entire body, but of the individual limbs as well. A parallel mission of this system is to provide the user with feedback regarding the quality of the movement performed, by attempting to match to the highest possible level the action executed to the strictly constructed prototype.

In this sense, our effort, apart from creating a system capable of handling actions from Martial Arts, aims to reveal the high level of scientific interest their study holds and to encourage more research towards the machine-based automated recognition of this class. Apart from the recognition though, efforts must be equally focused on the computer-assisted learning aspect of this problem, like the educational purpose of the feedback given by our system. Thus, the machine score can help karate students and teachers evaluate performance in a semi-automatic way and open the way to other user-interactive applications of similar recognition systems.

Related Work: Research efforts focusing on similar systems or at least partially, setting some groundwork for our endeavor, have made their appearance recently. Papers making use of local spatio-temporal features, like those produced by Histograms of Oriented Gradient(HOG) [1] or Histograms of Optical Flow(HOF) [2], extracted from regions surrounding detected spatial [3]or spatio-temporal [4]interest points, in order to train classification systems have brought attention to the high descriptive power of this combination. However, they have been used mainly for images [5, 6] or videos [7] of medium complexity and relatively short time duration, while in almost all cases in cooperation with Support Vector Machines (SVM) for the classification [8, 9]. Our approach will pursue to address description problems of longer and relatively more complex actions that comprise several stages.

Other previous works attempting to classify more advanced actions have used the widely known Hidden Markov Models (HMM) to create trainable prototypes, that work well for the representation of sequences with a more vast time depth. More specifically, in the area of studying movements from Martial Arts, there have been few attempts to create recognition systems [10, 11, 12] for such sequences, with the representation focusing in the description of specific body parts though, mainly the hands, while the features used were usually unable to capture the complexity of the problem. Our attempt focuses on bridging that gap, using a combination of highly descriptive local spatio-temporal features from Histographic methods and dynamically more potent prototypes created with HMMs.

2. PROPOSED SYSTEM

The system we introduce, as demonstrated in Figure 1, is composed by four discrete parts. At first, the input videos are processed in order to detect time-space interest points. These are used to construct 3D neighborhoods, from which we can extract our local spatio-temporal features using Histographic methods. In continuation, the features are properly transformed to decrease the complexity and improve the response time of the system. Finally, our transformed features are

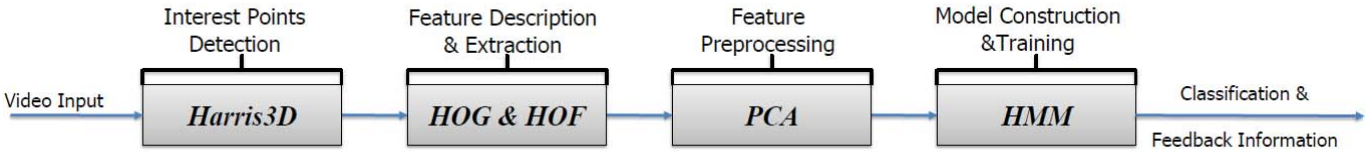


Fig. 1. Our Martial Arts video processing and recognition system.

inserted into our previously trained HMMs to achieve classification, while feedback regarding the similarity to the prototype is provided to the user. Details of the above methods are described below.

Feature Detection: The frame sequences of the input videos are processed by the Harris3D algorithm, proposed by Laptev and Lindeberg [4] for the extraction of Space-Time Interest Points (STIP). A linear spatio-temporal representation $D : \Re^2 \times \Re \rightarrow \Re^2_+$ of our image $f : \Re^2 \times \Re \rightarrow \Re$, with distinct spatial and temporal variances σ and τ, $\sigma_i^2 = s\sigma_l^2$, $\tau_i^2 = s\tau_l^2$, is:

$$D(\cdot; \sigma_l^2, \tau_l^2) = g(\cdot; \sigma_l^2, \tau_l^2) * f(\cdot) \quad (1)$$

where g is the spatio-temporal Gaussian kernel:

$$g(x, y, t; \sigma_l^2, \tau_l^2) = \frac{\exp(-(x^2 + y^2)/2\sigma_l^2 - t^2/2\tau_l^2)}{\sqrt{(2\pi)^3 \sigma_l^4, \tau_l^2}} \quad (2)$$

The second moment matrix is defined as:

$$\mu = g(\cdot; \sigma_i^2, \tau_i^2) * \begin{pmatrix} D_x^2 & D_x D_y & D_x D_t \\ D_x D_y & D_y^2 & D_y D_t \\ D_x D_t & D_y D_t & D_t^2 \end{pmatrix} \quad (3)$$

where $D_\xi(\cdot; \sigma_l^2, \tau_l^2) = \partial_\xi(g * f)$ are derivatives regularized via convolution with the Gaussian g.

The final 3D int.points are given by local maxima of the 3D Harris measure:

$$H = \det(\mu) - k \cdot \text{trace}^3(\mu) = \lambda_1\lambda_2\lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3 \quad (4)$$

where $\lambda_1, \lambda_2, \lambda_3$ are the eigenvalues of μ, but in a more simplified form where the best combination is selected between different space-time scales. The selected parameters are: $k = 0.0005$, $\sigma^2 = 4, 8, 16, 32, 64, 128$ and $\tau^2 = 2, 4$.

Feature Description & Extraction: After the detection of 3D interest points, we use the feature description introduced by Laptev et al. [7], extracting our features from the combined Histograms of Oriented Gradients (HOG) and Histograms of oriented Optical Flow (HOF) [1, 2] in the neighborhoods of the interest points, as can be seen in Figure2. The descriptor size is defined spatio-temporally by $\Delta_x(\sigma) = \Delta_y(\sigma) = 18\sigma$, $\Delta_t(\tau) = 8\tau$, while each volume is separated in a $n_x \times n_y \times n_t$ grid of cells. For each cell, 4-bin histograms of gradient orientations (HOG) and 5-bin histograms of optical flow (HOF) are computed, depicting local appearance and motion respectively, and after normalization they are concatenated into our feature vector. For our application, we used the grid parameters $n_x, n_y = 3, n_t = 2$. As a result, each interest point results to the computation of $3 \times 3 \times 2 \times 4 = 72$ HOG and $3 \times 3 \times 2 \times 5 = 90$ HOF features, meaning a feature vector of dimension: 3 (interest point spatio-temporal location - x,y,t) + 2 (spatio-temporal scales - σ^2, τ^2) + 72 + 90 = 167.

Feature Preprocessing: Due to the large dimension of our feature vectors, we apply Principal Components Analysis (PCA) on our HOG/HOF features, in order to select the coefficients that represent the problem without great loss of information. Using eigenvalue decomposition of our data matrix X, we can project it onto the reduced space defined by the first L unitary eigenvectors, $W_L : Y = W_L^T X$.

In our approach, we consider our feature vectors from consecutively detected 3D interest points as our repetitions and apply PCA on our 72 HOG and 90 HOF features separately, while maintaining the first 5 features (location-scales) intact. Thus, our data matrix is formed as:$[(samples)(Int.Points/sample) \times HOG/HOF]$ while we choose the first L eigenvectors that contain the **90%** of the variance-information of our data.

Models: As a final component of our system, we select Hidden Markov Models (HMMs) for the creation of our action prototypes, based on their ability to represent efficiently action types with multiple stages. We use linear left-to-right models with a mixture of different Gaussians for each state. The number of states and Gaussians varies between models of different movements, but specifically for Martial Arts we suggest models of 3-7 states, due to the complexity of the specific category, and 3-4 Gaussians per state, in order to have some flexibility, but maintain the strictness of our prototype.

After training, our models use the Maximum Likelihood algorithm to determine the most probable prototype to have produced the input sequence. Using our system as a Martial Arts training supplement, we extract the log probability for each video input and calibrate into a scale of 1-10 the similarity with the prototype through various scoring functions, providing feedback to the user.

3. EXPERIMENTAL PROCEDURE

For evaluating our system, we chose to apply it on Shotokan karate, a widespread but additionally very strict, regarding the athletes' technique and movement, form of Martial Arts.

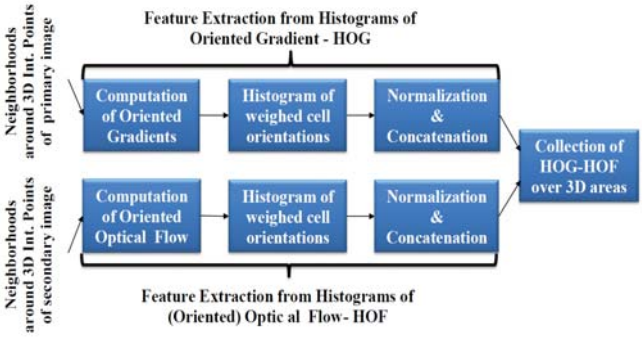
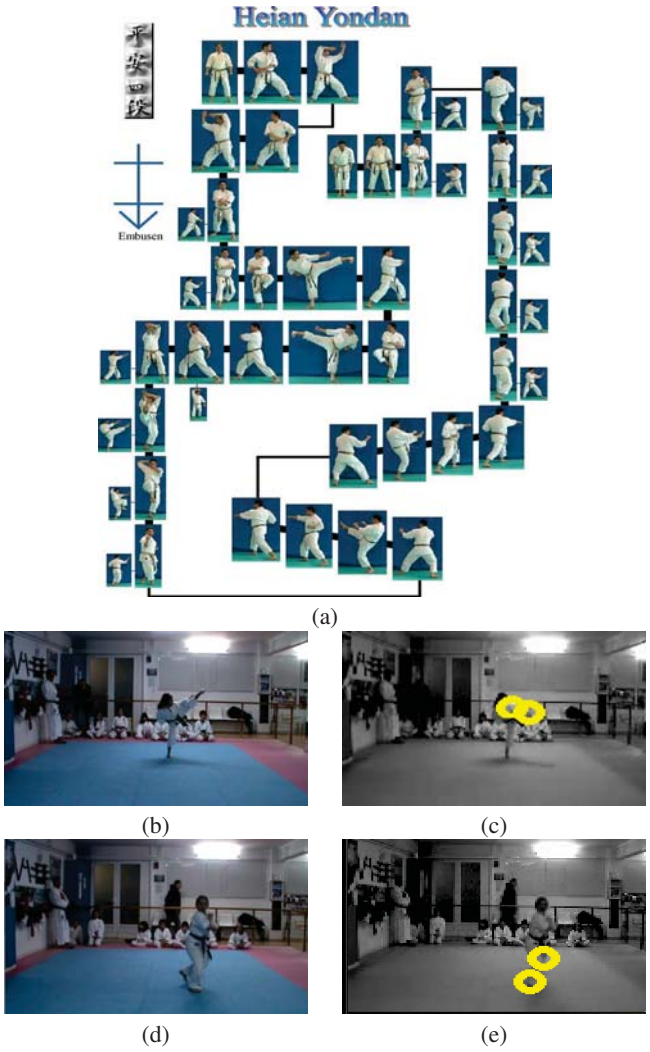


Fig. 2. Feature description using Histographic methods. HOG/HOF features extracted by computing orientation Histograms from cells in areas around 3D Int. points, after normalizing and concatenating.

Database: A new database of Shotokan karate movements (*katas*) was created, by filming karate athletes (*karatekas*) performing the 5 most basic kata sequences 'heian shodan', 'heian nidan', 'heian sandan', 'heian yondan' and 'heian godan'. In particular, we selected children athletes aged 7-14 for our experiment, in order to demonstrate how this system could be used as a training tool as well. The filming included 15 different karatekas and was carried out with medium resolution non-professional cameras, to demonstrate the flexibility of the system. An average of 25 different repetitions per kata were recorded, while all videos have an average duration of 1'30"-2'. The fixed camera was shooting from the side the athletes were facing the camera in the beginning and end of their movements.



(a) (b) (c) (d) (e)

Fig. 3. (a)Sequence of karate movement (kata) *heian yondan*. (b),(d) Frames from original video. (c),(e) Int. points detected on karateka.

Although the camera covers the area of the athletes' movement, there are occasions where athletes go out of our frame for some instants. Moreover, we have pursued the background to remain as uniform and still as possible, with no 'optical noise' meaning people and objects moving apart from our athletes, but in various samples that cannot be avoided. All the above increase the difficulty factor of recognition and classification of our database, since interest points may be detected in false positions as shown in Figure4, but simulate better the actual problem we are approaching. Finally, in order to evaluate the feedback our system provides to users, we have asked from the Shotokan karate instructor - *Sensei* to score the athletes movements, in order to compare them with our scoring functions.

Implementation: Working on a Linux/Ubuntu OS, we have converted our videos to 120x160 '.avi' of 25frames/sec. Through the application of Harris 3D, we detect 500-3000 interest points per sample, depending on its duration, receiving a $[Int.Points\times\ 167]$ matrix from the features extracted from the areas surrounding them. Afterwards, by applying PCA, using *Matlab*, on the 72 HOG and 90 HOF features separately, while keeping 90% of the variance, we manage to reduce the vector from 72 to 24 for the HOG and from 90 to 23 for the HOF features. In consequence, we form 4 categories of features to experiment with for the following stage. Features with no application of PCA (*nopca*) with $[Int.Points \times 167]$, features where PCA is only applied to HOG (*pcahog*) with $5 + 24 + 90 = 119$ resulting in $[Int.Points \times 119]$, features where PCA is only applied to HOF (*pcahof*) with $5 + 72 + 23 = 100$ resulting in $[Int.Points \times 100]$ and features where PCA is applied to HOG and HOF (*pca*) with $5 + 24 + 23 = 52$ resulting in $[Int.Points \times 52]$.

In continuation, we create our HMMs with the assistance of the *HTK* tool, converting our 2D data matrices into the appropriate format. Training is performed by using 4 or 5 samples that have received the highest scores from Sensei, 8-9-10, in order to build relatively strict prototypes, approximating the ideally performed kata. The rest of the samples of each kata category are going to be used for testing. For our experiments, we search through the best combinations among HMMs with 3-7 states and 1-4 mixtures of Gaussians per state for our 4 feature categories.

4. RESULTS

Through our testing procedure, we initially examine for 4 or 5 training samples, combinations of HMMs with 3-7 states $\times$ 1-4 Gaussians/state = 20 , while assuming the same model for all 5 kata types and for all 4 feature categories. The results can be seen in Figure5, where the overall accuracies are presented for each combination and we find that, conclusively, *nopca* features give us the best accuracy, followed by *pcahog*, *pca* and lastly, *pcahof*, while finest results through this search are 85-86%. In general, we also observe the increase of accuracy with the use of 5 training samples.

Higher precision can be attained once we try combinations of different HMM types for each kata by selecting more suitable model types according to the differences between different katas. In this way, we can retrieve combinations that result in accuracies of **88-89%** for the *nopca* feature category, but for the *pcahog* category as well. Particularly, with the *pcahog* of 4 training samples we manage to achieve overall accuracy of 88,54% for the combination [heianshodan, heiannidan, heiansandan, heianyondan, heiangodan]=[7s4g,

(a) (b) (c)

Fig. 4. Int. points detection on kata executions from our database. (a) Correct detection on karateka. (b) Detected also in false locations. (c) False detection.

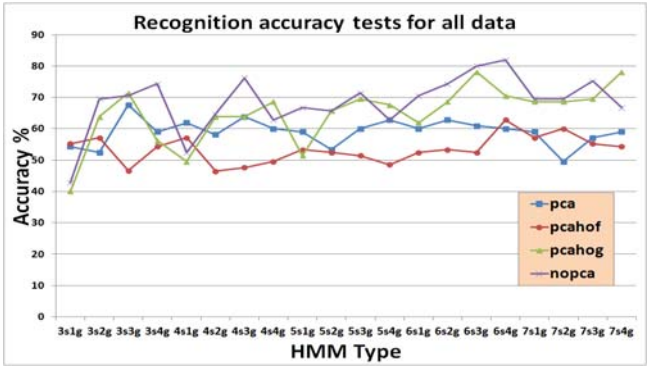


Fig. 5. Comparative plots of accuracies for different HMM combinations over all categories of features using 4 samples during training. The HMMs are the same for all 5 kata types.

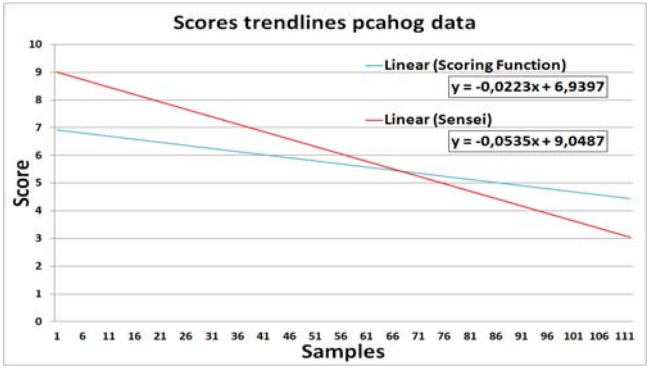


Fig. 7. Least mean squares trendlines of the scoring function compared to the 'Sensei' scoring for *pcahog* features using 4 training samples, scoring curve follows the slope of the 'Sensei' curve.

6s4g, 6s4g, 6s4g , 6s4g] (s states, g Gaussians/state), while the confusion matrix for this case can be seen in Figure6.

	shodan	nidan	sandan	yondan	godan
shodan	21	5	0	0	0
nidan	2	24	0	0	0
sandan	0	1	17	0	2
yondan	0	1	0	18	1
godan	0	0	0	0	13

(a)

	shodan	nidan	sandan	yondan	godan
shodan	22	3	0	0	0
nidan	4	21	0	0	0
sandan	0	0	15	1	3
yondan	0	1	0	18	0
godan	0	0	0	0	12

(b)

Fig. 6. Confusion matrices for the *pcahog* category of features using (a) 4 and (b) 5 samples during training.

Finally, using the log probability for each kata repetition we extract our feedback over the user's performance. In most cases, the scores given by our system coincide with the scores of the Sensei, but most importantly, the distinction between better and worse kata performances is similar. This is more easily observable if we create the trendlines for the scores in diminishing order using least mean squares interpolation, as for the scoring function

$$score(x) = 10\frac{logprob|x| - max\left[logprob|x|\right]/10}{max\left[logprob|x|\right] - max\left[logprob|x|\right]/10} \quad (5)$$

in Figure7, where the slope of our scoring function is quite similar to that of the Sensei. However, we observe that the result is worse for 5 training samples, which is rather expected since this way our prototype becomes less strict.

5. CONCLUSIONS

With this paper, we have introduced a novel Human Action Recognition system, mainly adjusted for the dynamic, multi-stage movements performed in Martial Arts, that has in its core the combination of HOG/HOF local spatio-temporal features extracted from neighborhoods of 3D interest points with the use of HMMs for our actions' prototypes. The system has been tested on our newly created database of Shotokan karate katas, resulting in relatively high accuracies reaching 89% and proving the strength of the above combination. In parallel, our system offers feedback to the user with similar to the authority discriminative attitude between better and worse executions with promising results for future optimization. Therefore, this action class provides fertile grounds for further research not only as a recognition problem, but regarding its interactive educational character as well.

Acknowledgements: Our deepest gratitude goes to the Sensei I. Stylianopoulos & the young Shotokan karate 'Yubukai Dojo' karatekas for the video footage.

6. REFERENCES

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*. 2005, IEEE Computer Society.

[2] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," *ECCV*, 2006.

[3] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE trans. PAMI*, 2005.

[4] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2, pp. 107–123, 2005.

[5] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. PAMI*, pp. 1627–1645, 2009.

[6] I. Laptev, "Improving object detection with boosted histograms," *Image and Vision Comp.*, vol. 27, no. 5, 2009.

[7] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, 2008.

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

[9] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *British Machine Vision Conf. on*, 2009.

[10] D.A. Becker, *Sensei: a real-time recognition, feedback and training system for t'ai chi gestures*, Ph.D. thesis, 1997.

[11] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden markov models for complex action recognition," in *CVPR*, 1997.

[12] N.M. Oliver, B. Rosario, and A. Pentland, "Graphical models for recognizing human interactions," *Advances in Neural Information Processing Systems*, pp. 924–930, 1999.