2-2015

Person Identification from Streaming Surveillance Video using Mid-Level Features from Joint Action-Pose Distribution

Binu M. Nair
University of Dayton, nairb1@udayton.edu

Vijayan K. Asari
University of Dayton, vasari1@udayton.edu

eCommons Citation

Nair, Binu M. and Asari, Vijayan K., "Person Identification from Streaming Surveillance Video using Mid-Level Features from Joint Action-Pose Distribution" (2015). *Electrical and Computer Engineering Faculty Publications*. 381.
https://ecommons.udayton.edu/ece_fac_pub/381

Person identification from streaming surveillance video using mid-level features from joint action-pose distribution

Binu M. Nair and Vijayan K. Asari

ECE Department, University of Dayton, 300 College Park, Dayton, OH-45469

ABSTRACT

We propose a real time person identification algorithm for surveillance based scenarios from low-resolution streaming video, based on mid-level features extracted from the joint distribution of various types of human actions and human poses. The proposed algorithm uses the combination of an auto-encoder based action association framework which produces per-frame probability estimates of the action being performed, and a pose recognition framework which gives per-frame body part locations. The main focus in this manuscript is to effectively combine these per-frame action probability estimates and pose trajectories from a short temporal window to obtain mid-level features. We demonstrate that these mid-level features captures the variation in the action performed with respect to an individual and can be used to distinguish one person from the next. Preliminary analysis on the KTH action dataset where each sequence is annotated with a specific person and a specific action is provided and shows some interesting results which verify this concept.

Keywords: Auto-Encoder, Action Recognition, Restricted Boltzmann Machines, Histogram of Flow, Local Binary Flow Patterns, Pose Trajectories

1. INTRODUCTION

Recent research in human motion analysis for surveillance applications in the last decade or so has been shifted more towards extracting features for human activity recognition, crowd flow analysis and pedestrian tracking. However, analyzing motion for the purpose of identifying individuals from surveillance has not been extensively researched on and has been limited to evaluating gait of individuals for person identification under controlled conditions such as in clinical studies. Moreover, the action in context has always been restricted to the walk action and study of motion patterns for any generic action from surveillance footage has not been the focus in the research community and industry. We address some of the questions with regards to identifying individuals: Does knowing the human action play an important part in extracting gait features for person identification? Can human action be an important cue to recognizing an individual other than just gait patterns? Intuitively, every person does the same action differently i.e the way in which a person does a certain action play an important part in recognizing that individual.

In this manuscript, we aggregate per-frame action probability estimates and pose trajectory estimates over a temporal window of length L frames and use a restricted Boltzmann machine (RBM) to capture the mid-level feature vector in real time. This feature vector is related to a local temporal window and can be obtained from a streaming video and therefore, can be used to identify an individual in a short period of time. Section 2 gives a survey of pose estimation and action recognition frameworks and shows that not much research has been done in evaluating action and pose to determine the identity of the individual. Section 3 provides the detailed description of the action association framework using auto-encoders, the pose trajectory latent features and the computation of the mid-level features using restricted Boltzmann machines. Finally, we provide our analysis of the algorithm in Section 4 with conclusions in Section 5.

Further author information : (Send correspondance to Binu M Nair
Binu M Nair: E-mail: nairb1@udayton.edu

Video Surveillance and Transportation Imaging Applications 2015, edited by Robert P. Loce, Eli Saber,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 9407, 94070P · © 2015 SPIE-IS&T
CCC code: 0277-786X/15/$18 · doi: 10.1117/12.2083423

2. LITERATURE SURVEY

In this section, we focus on some of recent state of the art techniques used in the field of human action recognition and pose estimation. Research in human action recognition can be grouped into different categories such as image models, sparse features and grammars/templates.[1] Some state of the art algorithms under the banner of image models used the concept of space time shapes and computed properties characterizing these shapes to classify actions.[2] However, the requirement of a good foreground segmentation for space time shapes rendered this approach impractical. To capture the temporal nature of an action, some earlier algorithms modeled features across time using Hidden Markov Models and Conditional Random Fields but this required predefined states for initialization of the model and sequence length normalization. By detecting sparse features or interest points on video sequences, a robust representation can be obtained by computing a histogram of these sparse features in a bag of words model.[3] Following this paradigm, a well-known interest point detector known as the Spatio-Temporal Interest Point (STIP) was proposed by Laptev et al.[4] This was extended detecting interesting events in video sequences and was extended to classify actions.[5] Each STIP is described by a HOG/HOF descriptor computed from its local neighborhood. Recent progress in human action and activity recognition used these STIP points and the Bag of Words model as low-level features in complex learning frameworks. Wang et al[6] developed a contextual descriptor for each STIP which described its local neighborhood and the arrangement of points using a probabilistic model. Yuan et al[7] computed the 3D-R transform to get the global arrangement of the STIP points and proposed a contextual SVM kernel for sequence classification. However, these algorithms focused mainly on automatic video annotation and not necessarily designed for real time association of an action (soft classification). For our purposes, we require action probability estimates computed at every frame and where the computaton of these estimates do not depend on the motion speed variance between individuals. The above mentioned methods based on STIP features are dependent on the number of frames of the sequence used in the computation of the histogram of codewords.

In recent years, the problem of human body pose estimation has not just being limited to tracking points or corners or using depth information. One of the state of art methods for human pose estimation on static images is the Flexible mixture of parts model, proposed by Yang and Ramanan.[8] Instead of explicitly using variety of oriented body part templates(parametrized by pixel location and orientation) in a search-based template matching scheme, a family of affine-warped templates is modeled, each template containing a mixture of non-oriented pictorial structures. Ramakrishna et al[9] proposed an occlusion aware algorithm which tracks human body pose in a sequence where the human body is modeled as a combination of single parts such as the head and neck and symmetric part pairs such as the shoulders, knees and feet. Here, the important aspect in this algorithm is that it can differentiate between similar looking parts such as the left or right leg/arm, thereby giving a suitable estimate of the human pose. In our algorithm, we use the flexible mixture of parts model in getting a human pose estimate at each frame.

3. PROPOSED METHODOLOGY

Here, we propose an algorithm to compute mid-level features using energy models from the joint distribution of action and pose trajectory features, accumulated over a temporal window of length L frames. The action features are conditional probability estimates of an action computed per frame by an auto-encoder based action classification scheme while the pose trajectory features are the body joint displacement estimates inferred from the per-frame pose estimates of the articulated parts model.[8] The action classification framework makes use of the optical-flow based motion features which was published in our action recognition algorithm[10] and the pose trajectory features are obtained by applying the trajectory descriptor[11,12] on the pose estimates. These pose features and action features are fed to a Restricted Boltzmann machine (RBM) to compute the mid-level features. The RBM structure after appropriate training, describes the joint probability of the action and pose features and reveals a hidden set of units which provides us with the mid-level features. These mid-level features are considered to be unique with respect to each individual and is fed to an SVM[13] based classification framework for person identification. A block schematic of the overall process is given in Figure 1.

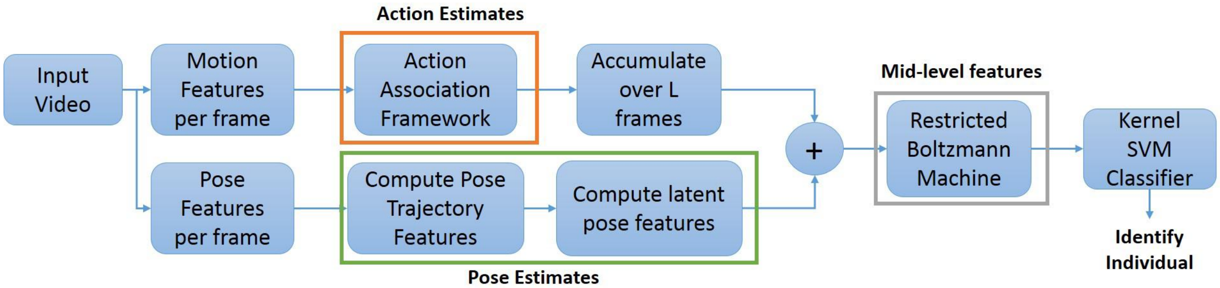


Figure 1: Block schematic of proposed algorithm.

3.1 Feature Extraction : Motion/Pose features

As shown in the block schematic, the motion and pose features are computed per frame. The motion features are extracted from the optical flow computed between two consecutive frames. The features are concatenation of the Hierarchical Histogram of Optical flow (HHOF),[10] Local Binary Flow patterns (LBFP)[10] and R-Transform (RT).[10] The HHOF gives the first order local and global information of the motion, the LBFP provides a second order information while the R-Transform provides a shape profile to the motion flow. These features have been proven that they are effective in providing action probability estimates in an action classification framework.[14, 15]

The pose features are the articulated pose model estimates at each frame where the location of 12 relevant joints namely left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle and right ankle are extracted. These joints approximates the distinct pose of an individual and its variation across a time is different for each individual when considering a specific action. Thus, the variation of these pose features as the person performing a specific action differs with each individual and can be an important cue in distinguishing individuals.

3.2 Low-level features

We can obtain low-level features for person identification by accumulating action discriminatory features and latent pose trajectory features. Action discriminatory features are per-frame conditional probability estimates obtained from an action classification framework and accumulated across a temporal window of length L frames. The pose trajectory features are normalized displacement vectors computed for each relevant body joint across the same temporal window. The latent pose trajectory features can be obtained by using a non-linear dimensionality reduction. Here, we use the auto-encoder for this purpose. Thus, low-level action discriminatory features and latent pose trajectory features can be used to identify an individual in a L frame temporal window.

3.2.1 Action Discriminatory features

We employ the use of auto-encoders to obtain action-specific temporal space for recognizing the action from per-frame motion features. For each action class, a trained auto-encoder which is a generalization of non-linear PCA, transforms the motion feature space into a time-independent space where the latent features lie on a temporal manifold. So, if the per-frame motion features of a test sequence with the true action class label 'walk' is applied to the trained auto-encoders of each action class, then the auto-encoder of the walk action class would give the least re-construction error. Similar are the cases of test sequences coming from other actions. This error forms the basis of determining the action class of that sequence as well as getting a probable estimate of the closeness of that sequence to the estimated class. An illustration of the concept is shown in Figure 2.

The auto-encoder is a neural network which obtains a lower-dimensional representation (encoder) and reconstructs the input features (decoder) using tied weights. This network is trained by treating a pair of layers as an RBM model and then using back-propagation to fine tune the weights. Once the network is trained, any motion feature at a frame can be encoded first and then reconstructed. If that motion feature does not belong to that specific auto-encoder, then the reconstruction error is large. This error can be interpreted as a conditional probability of the motion feature $\mathbf{x}_n$ belonging to the action class m. This is given in Equation 1 where $\chi^2(.)$ is

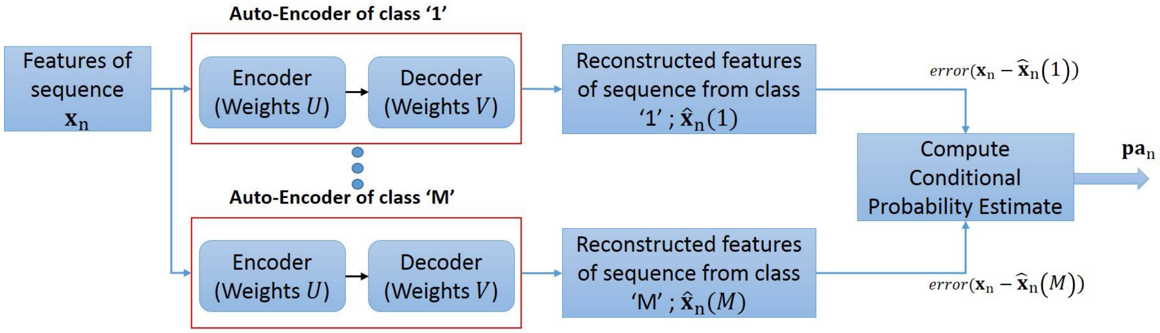

Auto-Encoder of class '1'
Features of sequence $\mathbf{x}_n$
Encoder (Weights U)
Decoder (Weights V)
Reconstructed features of sequence from class '1' ; $\hat{\mathbf{x}}_n(1)$
$error(\mathbf{x}_n - \hat{\mathbf{x}}_n(1))$
Compute Conditional Probability Estimate
$\mathbf{pa}_n$
Auto-Encoder of class 'M'
Encoder (Weights U)
Decoder (Weights V)
Reconstructed features of sequence from class 'M' ; $\hat{\mathbf{x}}_n(M)$
$error(\mathbf{x}_n - \hat{\mathbf{x}}_n(M))$

Figure 2: Block schematic of the action association framework.

the chi-squared metric, $\hat{\mathbf{x}}_n(m)$ is the reconstructed motion feature $\mathbf{x}_n$ obtained from the auto-encoder of action class m.

$$Prob(\mathbf{x}_n|class = m) = K\exp(-(\frac{\chi^2(\mathbf{x}_n - \hat{\mathbf{x}}_n(m))}{A_m})) \qquad (1)$$

Therefore, from a temporal window of length L frames and with M different action classes, the discriminatory action features can be summarized as $\mathbf{a}_{n,L} = [\mathbf{pa}_n^T\, \mathbf{pa}_{n+1}^T\, \mathbf{pa}_{n+2}^T....\mathbf{pa}_{n+L}^T] \in \mathbb{R}^{L.M}$ where $\mathbf{pa}_n = [pa_n^1.....pa_n^M]^T$.

3.2.2 Pose trajectory features

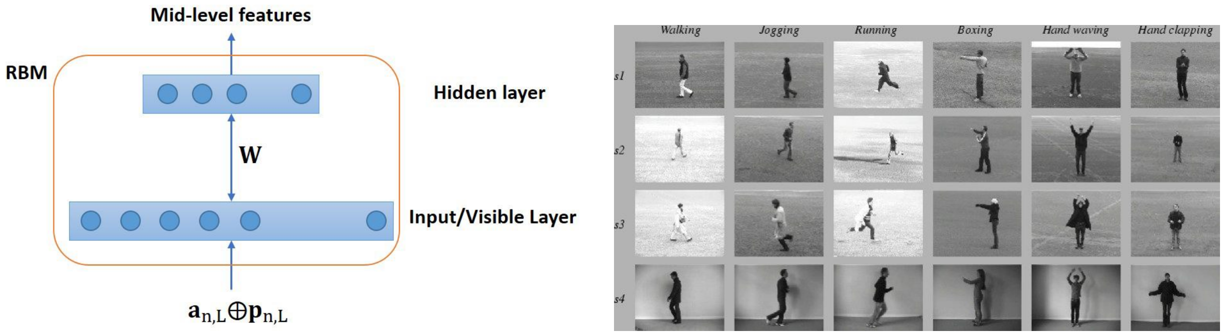
To compute the variation of the body joint across a temporal window of length L frames, we compute displacements between frames and normalize them. These are known as trajectory descriptors[12] and has been proved in our earlier published work[11] that it provides discrimination in recognizing specific gait patterns. The pose features are given as (x, y) locations of the relevant body joints. So, if there P body joints with their locations (x_n^p, y_n^p) at frame n, the displacement vector across L in the temporal window is given by $\mathbf{P}_{n,L} = [\mathbf{d}_n^T, \mathbf{d}_{n+1}^T, ...\mathbf{d}_{n+L}^T]^T$ where $\mathbf{d}_n = [(x_n^1 - x_{n-1}^1), (y_n^1 - y_{n-1}^1), (x_n^2 - x_{n-1}^2), (y_n^2 - y_{n-1}^2).........(x_n^P - x_{n-1}^P), (y_n^P - y_{n-1}^P)]^T$ is the displacement vector of the pose features. Then, we apply another trained auto-encoder to provide a lower-dimensional representation of the features, thereby obtaining the latent pose trajectory features $\mathbf{p}_{n,L}$.

3.3 Mid-Level features

The action discriminatory features $\mathbf{a}_{n,L}$ and the pose trajectory features $\mathbf{p}_{n,L}$ are computed for a temporal window of length L frames and this temporal window is scanned through the entire sequence. A shift of $L - 1$ frames corresponds to a streaming sequence. The combination action-pose features is obtained by concatenating them to form a single discriminatory feature $\mathbf{y}_{n,L} = \mathbf{a}_{n,L} \bigoplus \mathbf{p}_{n,L}$. Now, to find the joint distribution of the action features and pose features from the training data, we use a 3-layer deep restricted boltzmann machines. This is illustrated in Figure 3a. The hidden units at successive layers provide a higher level representation of the distribution of the action and pose features. Once the RBM layers are trained, then we obtain mid-level features which are sampled from action-pose distribution.

A Restricted Boltzmann Machine (RBM) is a undirected graphical energy model with two layers, the visible and the hidden layer consisting of independent units with only bi-directional weights associated between the layers. The features $\mathbf{y}_{n,L}$ are applied at the visible layer and when optimal weights are present, the hidden units provide the latent features which can efficiently be reconstructed back into the original feature space. The training of the RBM corresponds to finding optimal weights using the contrastive divergence technique.[16]

A support vector machine (SVM) classifier is trained using the mid-level features obtained from the RBM structure to distinguish between different individuals. Different types of kernels can be used and we prefer using a Chi-Squared kernel SVM.[13]



(a) Mid-level Feature Computation (b) Illustration of KTH Dataset.

Figure 3: Restricted Boltzmann Machine for midlevel feature computation (left) and illustration of action dataset (right).

4. EXPERIMENTS AND RESULTS

We test and evaluate our algorithm on a well-known KTH action dataset[17] which contains six types of human actions such as boxing (a1), hand waving (a2), hand clapping (a3), jogging (a4), running (a5) and walking (a6) are performed several times by 25 subjects in four different scenarios: outdoors $s1$, outdoors with scale variation $s2$, outdoors with different clothes $s3$ and indoors $s4$ as illustrated in Figure 3b. There are total of 600 sequences of which, each one is divided into sub-sequences of approximately 4 secs in length, resulting in a total of 2391 sub-sequences. All sequences were taken over homogeneous backgrounds with a static camera with $25fps$ and these sequences are down-sampled to a spatial resolution of 160×120. For our analysis, we consider the sequences of all individuals from set 1 which do not contain any scale variations or large clothing variations. In this set, we train the system on the first 3 sub-sequences and test on the 4^{th} sub-sequence. Each sequence is associated with a specific individual performing a specific action.

For our analysis, we provide two kinds of results. One is the set of accuracies obtained in identifying 5 different people where all are performing the same action (given in Table 1). The other are the set of accuracies obtained in identifying 5 different people when they are performing 3 different actions (given in Table 2. From Table 1, we see that our proposed algorithm is able to obtain an average of 56% for boxing action, 52% for hand waving, 55% for hand clapping action, 41.2% for jogging action, 32.2% for running action and 39.3% for walking action. Here, the boxing , hand waving and hand clapping action have reasonable accuracies while the walking, running and jogging have low accuracies due to the low number of training samples. Note that all these accuracies for person identification are obtained by computing mid-level features from a temporal window of length $L = 10$ frames. Thus, these results show a lot of promise in identifying individuals from a streaming surveillance video where motion features corresponding to a specific type of action are classified. Now, when multiple types of actions are involved in a surveillance, the accuracy in identifying an individual gets dropped to around 40% on an average. This is because of the lack of data for accurate training of the auto-encoder based action association framework and the mid-level feature computation. However, the algorithm shows the potential in deriving mid-level features and having an RBM network to represent joint action-pose probability distribution. If large amounts of training data are available, the RBM network can be trained more accurately which should result in a better generalization of the true action-pose distribution using the RBM network, resulting in better accuracies for person identification.

5. CONCLUSIONS

We have proposed an algorithm to derive mid-level features from an action-pose distribution estimated by using a series of auto-encoders and Restricted Boltzmann machines, to identify individuals irrespective of the action being performed. Our algorithm takes in conditional probability estimates of the action being performed as an additional cue along with the pose trajectory features along a temporal window. The mid-level features correspond to not only how the person moves but also how the person performs the action. These two cues are vital in recognizing an individual in a streaming surveillance scenario. Our proposed algorithm with preliminary

Person/Action	a1	a2	a3	a4	a5	a6
Persons 1-5	59.22%	53.2%	67.65%	33.62%	37.70%	56.19%
Persons 6-10	55.35%	56.28%	54.39%	52.67%	36.95%	42.42%
Persons 11-15	57.95%	48.90%	48.88%	47.27%	28.07%	30%
Persons 16-20	55.51%	39.37%	56.27%	26.56%	25.72%	21.81%
Persons 21-25	55.20%	62.95%	49.83%	44.95%	31.92%	46.15%

Table 1: Person Identification accuracy results for a set of 5 individuals for each action using midlevel features for a temporal window of length $L = 10$ frames.

Person/Action	a1,a2,a6	a1,a3,a6	a1,a2,a4	a1,a3,a4	a1-a6
Persons 1-5	48.47%	53.59%	42.33%	49.16%	43.99%
Persons 6-10	48.64%	45.24%	47.33%	43.54%	42.16%
Persons 11-15	43.23%	40.59%	45.56%	42.92%	36%
Persons 16-20	34%	43.10%	35.81%	45.07%	34.44%
Persons 21-25	42.14%	46.58%	52.65%	47.66%	46.53%

Table 2: Person Identification accuracy results for a set of 5 individuals and 3 actions for a temporal window of length $L = 10$ frames

results has proved that this concept is feasible and that additional evaluation of this approach will lead to a better understanding on identifying specific individuals.

REFERENCES

[1] Weinland, D., Ronfard, R., and Boyer, E., "A survey of vision-based methods for action representation, segmentation and recognition," *Computer Vision and Image Understanding* **115**(2), 224 – 241 (2011).

[2] Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R., "Actions as space-time shapes," in [*Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*], **2**, 1395–1402 Vol. 2 (2005).

[3] Niebles, J., Wang, H., and Fei-Fei, L., "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision* **79**(3), 299–318 (2008).

[4] Laptev, I., "On space-time interest points," *International Journal of Computer Vision* **64**(2-3), 107–123 (2005).

[5] Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B., "Learning realistic human actions from movies," in [*Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*], 1–8 (2008).

[6] Wang, J., Chen, Z., and Wu, Y., "Action recognition with multiscale spatio-temporal contexts," in [*Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*], 3185–3192 (2011).

[7] Yuan, C., Li, X., Hu, W., Ling, H., and Maybank, S., "3d r transform on spatio-temporal interest points for action recognition," in [*Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*], 724–730 (2013).

[8] Yang, Y. and Ramanan, D., "Articulated pose estimation with flexible mixtures-of-parts," in [*Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*], 1385–1392 (June 2011).

[9] Ramakrishna, V., Kanade, T., and Sheikh, Y., "Tracking human pose by tracking symmetric parts," in [*Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*], 3728–3735 (2013).

[10] Nair, B. and Asari, V., "Learning and association of features for action recognition in streaming video," in [*Advances in Visual Computing*], Bebis, G., Boyle, R., Parvin, B., Koracin, D., McMahan, R., Jerald, J., Zhang, H., Drucker, S., Kambhamettu, C., El Choubassi, M., Deng, Z., and Carlson, M., eds., *Lecture Notes in Computer Science* **8888**, 642–651, Springer International Publishing (2014).

[11] Nair, B., Kendricks, K., Asari, V., and Tuttle, R., "Body joint tracking in low resolution video using region-based filtering," in [*Advances in Visual Computing*], Bebis, G., Boyle, R., Parvin, B., Koracin, D., McMahan, R., Jerald, J., Zhang, H., Drucker, S., Kambhamettu, C., El Choubassi, M., Deng, Z., and Carlson, M., eds., *Lecture Notes in Computer Science* **8887**, 619–628, Springer International Publishing (2014).

[12] Wang, H., Klser, A., Schmid, C., and Liu, C.-L., "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision* **103**(1), 60–79 (2013).

[13] Chang, C.-C. and Lin, C.-J., "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology* **2**, 27:1–27:27 (2011). Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[14] Nair, B. M. and Asari, V. K., "Time invariant gesture recognition by modelling body posture space," in [*Proceedings of the 25th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems: Advanced Research in Applied Artificial Intelligence*], *IEA/AIE'12*, 124–133, Springer-Verlag (2012).

[15] Nair, B. and Asari, V., "Regression based learning of human actions from video using hof-lbp flow patterns," in [*Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*], 4342–4347 (Oct 2013).

[16] Hinton, G. E. and Salakhutdinov, R. R., "Reducing the dimensionality of data with neural networks," **313**(5786), 504–507 (2006).

[17] Schuldt, C., Laptev, I., and Caputo, B., "Recognizing human actions: a local svm approach," in [*Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*], **3**, 32–36 Vol.3 (2004).