Fusion of Color and Depth Video for Human Behavior Recognition in an Assistive Environment

Dimitrios I. Kosmopoulos[1], Paul Doliotis[2], Vassilis Athitsos[2], and Ilias Maglogiannis[3]

[1] TEI of Crete, Dept of Applied Informatics and Multimedia, GR-71500, Greece
[2] University of Texas at Arlington, Dept of Computer Science and Engineering, TX-76013, USA
[3] Dept of Digital Systems, University of Piraeus, GR-18534
dkosmo@ieee.org, {doliotis,athitsos}@uta.edu, imaglo@unipi.gr

Abstract. In this paper we investigate the effects of fusing feature streams extracted from color and depth videos, aiming to monitor the actions of people in an assistive environment. The output of fused time-series classifiers is used to model and extract actions. To this end we compare the Hidden Markov model classifier and fusion methods like early, late or state fusion. Our experiments employ a public dataset, which was acquired indoors.

1 Introduction

One of the key questions in creating pervasive systems for the care of the elderly is the graceful integration with the human user [1]. Towards building such a system a highly desired property that needs to be satisfied is "non-intrusiveness". Computer vision methods can satisfy this property and are typically used in assistive environments. One of the main challenges is to transform the video stream into a useful source of information. This can be further divided in several sub-problems like how to track people in the captured video stream, how to recognize their postures and how to analyze their short term actions and long term behaviors.

Motion analysis in video, and particularly human behaviour understanding, has attracted many researchers [2], mainly because of its fundamental applications in video surveillance, video indexing, virtual reality and computer-human interfaces. The automatic modeling and recognition of human behaviour to reduce human intervention in assistive or other environments is one of the most challenging problems in computer vision. The related systems are envisaged to automatically detect, categorize and recognize human behaviors, calling for human attention only when necessary. This is expected to increase the effectiveness of 24/7 monitoring services for elderly or patients and make such services financially viable [3].

N. Streitz and C. Stephanidis (Eds.): DAPI/HCII 2013, LNCS 8028, pp. 42–51, 2013.

There are several works on human behavior recognition in assistive environments using color cameras, e.g., [4], [5]. The color information captured by conventional cameras is a very useful cue, which can be used for environment modeling and object tracking. Problems which are associated to color video tracking are the illumination changes as well as the occlusions [6]. Furthermore, since human motion is essentially three-dimensional, the information loss in the depth channel could cause degradation of the representation and discriminating capability for these feature representations. The emergence of affordable depth sensors (e.g., Microsoft Kinect) which are largely unaffected by illumination (at least indoors) has facilitated capturing in real-time not only color videos, but also depth videos with acceptable resolution (e.g., 640×480 in pixel) and accuracy (e.g., = 1cm). By employing appropriate methods we can extract three-dimensional and motion information of the monitored subjects in the scene. Therefore the depth ambiguity of the color camera could be bypassed. On the other hand such depth sensors cannot differentiate between objects of the same depth different color, which is trivial for color cameras.

Clearly the color and depth information are correlated but also complementary to a large extent, so it would be expected to have considerable benefits by fusing them appropriately together aiming at more robust pervasive behavior recognition systems. The contribution of this work is a study of the performance of fusion techniques that combine color and depth videos for human activity analysis. To this end we use the RGBD-HuDaAct dataset [7], which is publicly available. We compare fusion methods at the decision level, the feature level and the state level.

The rest of this paper is structured as follows. In the following section we briefly survey the related work regarding systems employing color and depth information. Section 3 describes the feature extraction and the fusion approaches that we employed. Section 4 describes the experimental results and finally section 5 concludes this paper.

2 Related Work

One of the earliest works on action recognition using a depth sensor was presented in [8]. In that paper the authors employ an action graph to model explicitly the dynamics of the actions and a bag of 3D points to characterize a set of salient postures that correspond to the nodes in the action graph. That method managed to halve the recognition errors comparing to the 2D silhouette based recognition. However one of the main limitations was that it completely ignored color information.

Ni et al.[7] proposed a method for human activity recognition that takes into account such color information coupled with depth information. They proposed two multimodality fusion schemes, which simply combine color and depth streams by concatenation and are developed from two state-of-the-art feature representation methods for action recognition, i.e., spatio-temporal in-terest points (STIPs) and motion history images (MHIs).

Depth and color data can also provide higher lever and more meaningfull features like skeletal joints of a person. Sung et al. [9] propose a supervised learning approach in which they collected ground-truth labeled data for training their model. Their input was color and depth images from a Kinect sensor, from which they extracted certain features (like skeletal joints) that were fed as input to a learning algorithm. They trained a two-layered maximum-entropy Markov model which captured different properties of human activities, including their hierarchical nature and the transitions between sub-activities over time.

However skeletal joint data aren't always available, especialy in scenarios, where the camera is mounted on the ceiling. Zhao et al. [10] addressed that issue and in their work they investigated performances of different ways of extracting interest points, since interest point based approaches can handle cluttered background and partial occlusions. Additionaly they proposed a local depth pattern to represent each local video volume at each interest point. They used LibSVM [11] to classify human activities in a multi-class fashion.

While Zhao et al. investigated performances of different ways of extracting interest points for activity recognition, in this paper we investigate performances of fusion techniques that fuse color and depth. In contrast to other methods we investigate fusion schemes at the state level of the popular HMM framework, which can give better results than the simple fusion schemes that rely on concatenation of the input feature streams. In this paper only region descriptors are used, however the fusion approach has no constraints regarding the type of the employed features.

3 Methodology

The proposed methodology performs initially a feature extraction step, combining the two different sources: depth videos and color videos. From the depth videos we extract two different types of feature vectors (forward and backward), as will be described next, while from the color video we calculate features describing the human blob. The features from the whole sequence are combined and given as input to a classifier, which in turn decides on the performed activity. The method is applicable on segmented actions, but can also be used for online classification, by integration with a particle filter that makes hypotheses about sequences of actions (see, e.g., [12]).

3.1 The Features

Features from Color Images. The image features that were extracted from color images were based on a variation of Motion History Images (MHIs). MHIs are among the first holistic representation methods for behavior recognition [13]. In an MHI H_τ, pixel intensity is a function of the temporal history of motion at that point. In [14], it was shown that pixel change history (PCH) images are able to capture relevant duration information with better discrimination performance.

The PCH of a pixel is defined as:

$$P_{\varsigma,\tau}(x,y,t) = \begin{cases} \min(P_{\varsigma,\tau}(x,y,t-1) + \frac{255}{\varsigma}, 255) \\ \text{if} D(x,y,t) = 1 \\ \max(P_{\varsigma,\tau}(x,y,t-1) - \frac{255}{\tau}, 0) \\ otherwise \end{cases} \quad (1)$$

where $P_{\varsigma,\tau}(x,y,t)$ is the PCH for a pixel at (x,y), $D(x,y,t)$ is the binary image indicating the foreground region, ς is an accumulation factor and τ is a decay factor. By setting appropriate values to ς and τ we are able to capture pixel-level changes over time. The result is a scalar-valued image where more recently moving pixels are brighter.

Assuming that the human blob shapes during specific actions have discriminative capabilities we use the complex Zernike moments to capture the PCH images, which provide scale invariant representations and are relatively robust to noise. The moments of order p are defined on an grayscale image f as:

$$A_{pq} = \frac{p+1}{\pi} \int_0^1 \int_{-\pi}^{\pi} R_{pq}(r) e^{-jq\theta} f(r,\theta) r dr d\theta \quad (2)$$

where $r = \sqrt{x^2+y^2}$, and $\theta = \tan^{-1}(\text{y/x})$ and $-1 < x, y < 1$ (x,y are the image coordinates, with respect to the center, around which the integration is calculated) and:

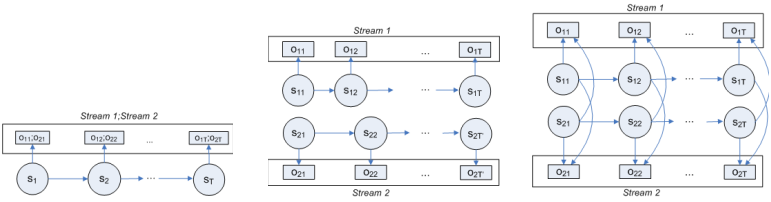
$$R_{pq}(r) = \sum_{s=0}^{\frac{p-q}{2}} (-1)^s \frac{(p-s)!}{s!(\frac{p+q}{2}-s)!(\frac{p-q}{2}-s)!} r^{p-2s} \quad (3)$$

where $p - q = even$ and $0 \leq q \leq p$. Moments of low order hold the coarse information while the ones of higher order hold the fine details. However, the more detailed the region representation is, the more processing power will be demanded, and thus a trade-off has to be reached considering the specific application requirements.

The MHI images are represented by means of the complex Zernike coefficients A_{00}, A_{11}, A_{20}, A_{22}, A_{31}, A_{33}, A_{40}, A_{42}, A_{44}, A_{51}, A_{53}, A_{55}, A_{60}, A_{62}, A_{64}, A_{66}, for each of which the norm and the angle were included in the provided descriptors. We used a total of 31 parameters (constant elements were removed), thus providing an acceptable scene reconstruction without a computationally prohibitive dimension.

Features from Depth Images. Ni et al. [7] proposed the use of a depth sensor and they introduced the motion history along the depth changing directions. To encode the backward motion history (decrease of depth), they introduced the backward-DMHI (bDMHI):

$$H_\tau^{bD}(x,y,t) = \begin{cases} \tau, \text{if } D(x,y,t) - D(x,y,t-1) < -\delta I_{th} \\ max(0, H_\tau^{bD}(x,y,t-1) - 1), \text{otherwise.} \end{cases} \quad (4)$$



(a) HMM with fused observations (b) Parallel HMM (PHMM) (c) Multistream fused HMM (MFHMM)

Fig. 1. Various fusion schemes using the HMM framework for two streams. The s, o stand for the states and the observations respectively. The first index marks the stream and the second the time.

(a) Color image (b) Backward MHI (from depth) (c) Forward MHI (from depth) (d) PCH image (from color)

Fig. 2. Illustration of MHI and PCH images for the *put on jacket* action

Here, H_t^{bD} denotes the backward motion history image and $D(x, y, t)$ denotes the depth sequence. δI_{th} is the threshold value for generating the mask for the region of backward motion.

Similarly, the forward history image, is defined as:

$$H_\tau^{fD}(x, y, t) = \begin{cases} \tau, \text{if } D(x, y, t) - D(x, y, t-1) > \delta I_{th} \\ max(0, H_\tau^{fD}(x, y, t-1) - 1), \text{otherwise.} \end{cases} \quad (5)$$

In order to calculate the depth change induced motion history images, according to the above equations, we use depth maps captured by a Kinect device. To tackle the problem of noise, we used a median filtering at the spatial domain. In the temporal domain each pixel value was replaced by the minimum of its neighbors. Similarly to the color images, each frame was represented by the 6-th order Zernike moments.

3.2 Fusion

As mentioned earlier, the depth and color images are highly complementary. Therefore, we can infer that by applying an appropriate fusion method we could achieve behavior recognition results better than the results that we could attain by using the information obtained by the individual data streams independently of each other. In the following, we shall survey the most popular fusion methods within the HMM framework and examine their applicability. The observations from this analysis will form the criterion for our selection of the most suitable HMM-based information fusion scheme to be used in the context of our system.

Existing approaches can be grouped into feature (or early) fusion and late fusion approaches. *Feature fusion* is the simplest approach; it assumes that the observation streams (sequences of feature vectors as defined in section 3.1) are synchronous. This synchronicity is a valid assumption for cameras that have overlapping fields of view and support synchronization. The related architecture *FHMM* is displayed in Fig. 1(a). Let us denote as s_t the FHMM state emitting the tth observation. Let us consider data deriving from a number of C observation streams, and denote as $\{\boldsymbol{o}_{1t}, \ldots, \boldsymbol{o}_{Ct}\}$ the observations at time t deriving from the available streams. Then, the full observation vector, $\boldsymbol{o}_t$, considered by the feature fusion approach at time t, is a simple concatenation of the available individual observations:

$$\boldsymbol{o}_t = \left(\boldsymbol{o}_{ct}^{'}\right)_{c=1\ldots C}^{'} \tag{6}$$

This way, the observation emission probability of the state $s_t = i$ of the fused model, when modeled as a k-component mixture model, yields:

$$P(\boldsymbol{o}_t|s_t = i) = \sum_{k=1}^{K} w_{ik} P(\boldsymbol{o}_t|\theta_{ik}) \tag{7}$$

where w_{ik} denotes the weights of the mixture components, and θ_{ik} are the parameters of the kth component density of the ith model state (e.g., mean and covariance matrix of a Gaussian pdf).

The major limitations of the feature fusion approach lie in the fact that the simple concatenation of observations from different streams leads to high dimensionality and often fails to capture significant statistical dependencies between the different sources of information.

An alternative that assumes that the observation streams are independent of each other is the *parallel HMM - PHMM* [15] (see Fig. 1(b)). This HMM-type model can be applied to cameras (or other sensors) that may not be synchronized and may operate at different acquisition rates. A PHMM does also comprise a number of component *streamwise* HMMs, independently trained of one another. Similar to the synchronous case, each stream c may have its own weight r_c depending on the reliability of the source. As a consequence of this construction, the PHMM suffers from the major disadvantage of tending to neglect any dependencies on the state level between the observation streams.

The *multistream fused HMM - MFHMM* is another method recently proposed for multistream data modeling [16] (see Fig. 1(c)). The connections between the component *streamwise* HMMs of this model are chosen based on a probabilistic fusion model, which is optimal according to the maximum entropy principle and a maximum mutual information criterion for selecting dimension-reduction transforms [16]. Specifically, if we consider a set of multistream observations $O = \{\boldsymbol{o}_t\}_{t=1}^{T}$, with $\boldsymbol{o}_t = \{\boldsymbol{o}_{ct}\}_{c=1}^{C}$, and $\boldsymbol{o}^c = \{\boldsymbol{o}_{ct}\}_{t=1}^{T}$, the MFHMM models this data based on the fundamental assumption

$$P(O) = \frac{1}{C}\sum_{c=1}^{C} P(\boldsymbol{o}^c)\prod_{r\neq c} P(\boldsymbol{o}^r|\hat{s}_c) \tag{8}$$

where $\hat{s}_c$ is the estimated hidden sequence of emitting states that corresponds to the cth stream observations, obtained by means of the Viterbi algorithm, $P(\boldsymbol{o}^c)$ is the observation probability of the cth stream observed sequence, and $P(\boldsymbol{o}^r|\hat{s}_c)$ is the *coupling density* of the observations from the rth stream with respect to the states of the cth stream model

$$P(\boldsymbol{o}^r|\hat{s}_c) = \prod_{t=1}^{T} P(\boldsymbol{o}_{rt}|\hat{s}_{ct}) \tag{9}$$

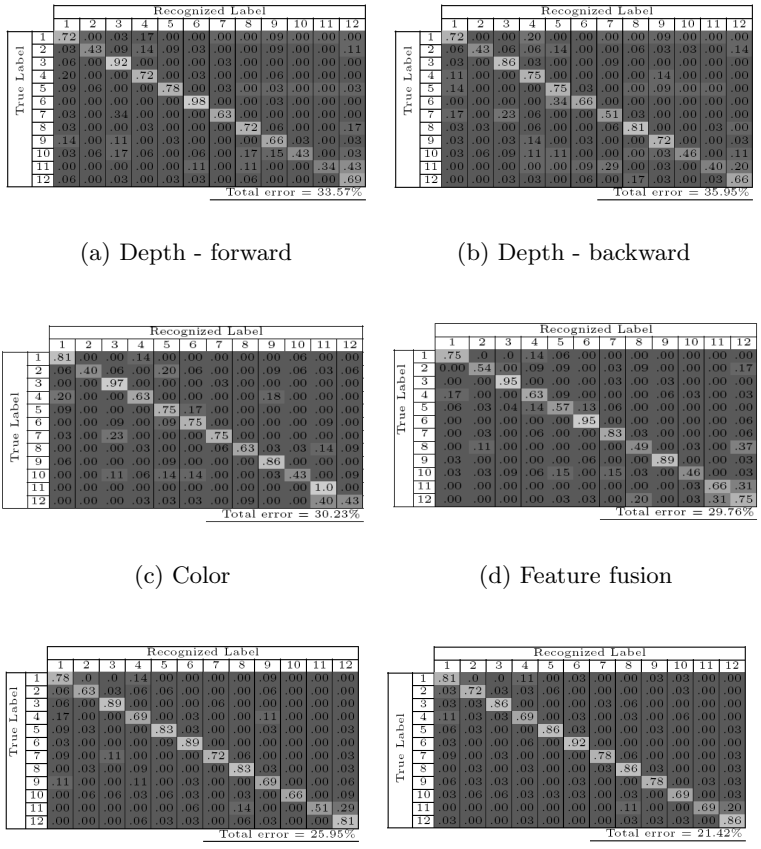
The probabilities $P(\boldsymbol{o}_{rt}|\hat{s}_{ct})$ of the MFHMM can be modeled by means of mixtures of Gaussian densities, similar to the state-conditional likelihoods of the *streamwise* HMMs.

Note also that for each possible value, say i, of $\hat{s}_{ct}$, i.e. for each different state of the streamwise HMMs, a different *coupling density model* $P(\boldsymbol{o}_{rt}|\hat{s}_{ct} = i)$ has to be postulated. Hence, if we consider K-state *streamwise* HMMs, there is a total of K different finite mixture models that must be trained to model the *coupling densities* $P(\boldsymbol{o}_{rt}|\hat{s}_{ct})$, $\forall r, c$.

4 Experiments and Results

The employed dataset is the RGBD-HuDaAct [7], which includes twelve categories of human daily activities, motivated by the definitions provided by health-care professionals. Namely these are: (1)*make a phone call*, (2)*mop the floor*, (3)*enter the room*, (4)*exit the room*, (5)*go to bed*, (6)*get up*, (7)*eat meal*,(8)*drink water*, (9)*sit down*, (10)*stand up*, (11)*take off the jacket* and (12)*put on the jacket*. There is also a category named as background activity that contains different types of random activities. Thirty actors were involved in capturing. The actors were student volunteers, and were asked to perform each activity 2 - 4 times. Finally, approximately 46 hours of video were acquired for a total of 1189 labeled video samples. Each video sample spans about 30 - 150 seconds.

The resolutions of both color image and depth map are 640×480 pixels. The color image is of 24-bit RGB values; each depth pixel is an 16-bit integer. Both



(a) Depth - forward (b) Depth - backward

(c) Color (d) Feature fusion

(e) Parallel HMM (f) Multistream HMM

Fig. 3. Confusion matrices for the twelve tasks in the RGBD-HuDaAct dataset. The results are normalized based on the total number of actions, considering all cross-validation runs. The actions are: (1)*make a phone call*, (2)*mop the floor*, (3)*enter the room*, (4)*exit the room*, (5)*go to bed*, (6)*get up*, (7)*eat meal*,(8)*drink water*, (9)*sit down*, (10)*stand up*, (11)*take off the jacket* and (12)*put on the jacket*.

sequences are synchronized and the frame rates are 30 frames per second. The color and depth frames are stereo-calibrated. The horizontal and vertical distances from the camera to the scene center under capture are about two meters each and the average depth of the human subject in the scene is about three meters (i.e., which is the optimal operation range of the depth camera). This geometric setting is appropriate for home or hospital ward monitoring.

The basic observations about the dataset have to do with the complementarity of the two sources of information: color images and depth. The latter is able to differentiate between actions that take place within the human blob, e.g., *make*

a phone call and *drink water* may look similar in color videos, however the depth motion is different. On the contrary depth sensors have problems when viewing objects with large discontinuities (e.g., actions *sit down*, *get up*, where furnitures are present); such depth maps have a significant amount of noise. After frame differencing and thresholding motion can be falsely detected even in areas where there are only still objects, while color cameras are much more robust concerning this aspect.

We have performed cross validation testing for one user after training for the rest ones. In all cases we used six-state continuous HMMs with two components for each state, which was gave reasonable results. The results are displayed in fig 3. Clearly the multi - stream approach outperforms the other methods, followed by the parallel HMM fusion. The feature fusion is clearly inferior and this is a result that agrees with the observations in [17], where a similar comparison was performed. The overall accuracies are close to the ones reported in [7], however the results are not directly comparable due to differences in the cross validation procedures. In [7] random sampling was performed to separate the training set from the test set, which was not replicated here. However, by establishing a fair comparison between the three fusion methods we were able to assess the early fusion scheme, which was the sole method tested in [7], in comparison to the other two fusion methods (parallel and multistream).

5 Conclusions

This paper investigated the effects of fusing color and depth videos, aiming to monitor the behavior of people in an assistive environment. The output of fused time-series classifiers was used to model and extract behaviors. To this end we employed the Hidden Markov model general framework. Fusion methods like early, late or state fusion were compared. The results from early fusion were weak compared to the other approaches. The late fusion gave better results, however the state fusion scheme outperformed all other methods. Our results are inline with the study in [17] for some different scenarios (industrial workflows). We expect that they can be generalized to other feature streams (e.g., spatiotemporal interest points) and we aim to investigate this hypothesis in the future.

Acknowledgment. This research has been co-financed by the European Union (European Social Fund ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: THALES. Investing in knowledge society through the European Social Fund.

References

1. Stanford, V.: Using pervasive computing to deliver elder care. IEEE Pervasive Computing 1(1), 10–13 (2002)

2. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Underst. 104(2), 90–126 (2006)
3. Doukas, C., Maglogiannis, I.: Emergency fall incidents detection in assisted living environments utilizing motion, sound, and visual perceptual components. IEEE Transactions on Inf. Techn. in Biomedicine 15(2), 277–289 (2011)
4. Antonakaki, P., Kosmopoulos, D., Perantonis, S.J.: Detecting abnormal human behaviour using multiple cameras. Signal Processing 89, 1723–1738 (2009)
5. Kosmopoulos, D.: Multiview behavior monitoring for assistive environments. Universal Access in the Information Society 10, 115–123 (2011)
6. Christodoulidis, A., Delibasis, K.K., Maglogiannis, I.: Near real-time human silhouette and movement detection in indoor environments using fixed cameras. In: PETRA 2012, pp. 1:1–1:7. ACM (2012)
7. Ni, B., Wang, G., Moulin, P.: Rgbd-hudaact: A color-depth video database for human daily activity recognition. In: ICCV Workshops, pp. 1147–1153 (2011)
8. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3d points. In: CVPR4HB 2010, pp. 9–14 (2010)
9. Sung, J., Ponce, C., Selman, B., Saxena, A.: Unstructured human activity detection from rgbd images. In: Int. Conf. Robotics and Automation, pp. 842–849 (2012)
10. Zhao, Y., Liu, Z., Yang, L., Cheng, H.: Combing rgb and depth map features for human activity recognition. In: 2012 Asia-Pacific Signal Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1–4 (December 2012)
11. Chang, C.C., Lin, C.J.: Libsvm: A library for support vector machines. ACM Trans. Intell. Syst. Technol. 2(3), 27:1–27:27 (2011)
12. Kosmopoulos, D., Doulamis, N., Voulodimos, A.: Bayesian filter based behavior recognition in workflows allowing for user feedback. Computer Vision and Image Understanding 116, 422–434 (2012)
13. Davis, J.W., Bobick, A.F.: The representation and recognition of action using temporal templates. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 928–934 (1997)
14. Xiang, T., Gong, S.: Beyond tracking: Modelling activity and understanding behaviour. Int. J. Comput. Vision 67(1), 21–51 (2006)
15. Chenand, C., Liang, J., Zhao, H., Hu, H., Tian, J.: Factorial HMM and parallel HMM for gait recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 39(1), 114–123 (2009)
16. Zeng, Z., Tu, J., Pianfetti, B., Huang, T.: Audio-visual affective expression recognition through multistream fused HMM. IEEE Transactions on Multimedia 10(4), 570–577 (2008)
17. Kosmopoulos, D., Chatzis, S.: Robust visual behavior recognition. IEEE Signal Processing Magazine 27(5), 34–45 (2010)