INTERACTION RECOGNITION IN WIDE AREAS USING AUDIOVISUAL SENSORS

Murtaza Taj and Andrea Cavallaro

School of Electronic Engineering and Computer Science
Queen Mary University of London, Mile End Road, London, E1 4NS, UK

ABSTRACT

We present an event recognition framework to detect interactions among objects, for example people, using a network of cameras and associated microphone pairs. The complementarity of the video and audio modalities is exploited to cover wide areas. In particular, object movements in portions of the scene that are not covered by the cameras' fields of view are estimated using the input from microphones. After estimating trajectories using audio-visual features, we recognize interactions based on a Coupled Hidden Markov Model Maximum a Posteriori (CHMM-MAP) approach. The states of the CHMM are initialized via Gaussian Mixture Model (GMM) clustering on a multi-dimensional feature space. Evaluation and comparison with three alternative methods demonstrate the effectiveness of the proposed CHMM-MAP trained on multiple features on both synthetic and real data.

***Index Terms*—** Event recognition, Audiovisual processing, Multimodality, Hidden Markov Model, GMM clustering

1. INTRODUCTION

The recognition of interactions is of great interest for surveillance, sports analysis and medical research. When scenes to be monitored are not covered completely by a single sensor, multiple cameras help observing behaviours. However, even multiple cameras may be insufficient to cover the whole scene, thus reducing the number of available object observations. Existing works [2] using multiple cameras for extended coverage are limited by their fields of view (FOV) and naturally fail when targets exit the observed regions. To overcome this problem, object positions can be estimated by prediction (based on the available observations in the cameras' FOV and their motion dynamics [7]) or by coupling cameras with sensors having a wider field of observation, such as microphones (Figure 1). When objects to be tracked emit sounds, the use of microphones also overcome some limitations of cameras, such as dealing with changing lighting conditions and visual occlusion due to vegetation or dust. The combination of audio and visual inputs allows us to improve position estimations that will in turn be used for interaction recognition.

In this paper we propose a framework for interaction recognition in multimodal sensor networks. This framework generates complete target tracks in complex scenes with multiple objects using audio-audio and audio-visual fusion, and operates without fixed event templates. After the estimation of the trajectories, the proposed algorithm extracts relative features for any pairs of interacting targets in the scene. Model training is performed on these relative features using a Coupled Hidden Markov model (CHMM) Maximum a Posteriori approach that allows us to obtain an improved modeling of full coupling between two processes by incorporating prior distribution. Interaction recognition is performed using Viterbi decoding on a set of events that includes various combinations of actions such as *approaching*, *meeting*, *going together*, *going separately* and *following*, which are of particular interest for analyzing behaviours in a scene. The flow diagram of the proposed framework is shown in Fig. 2.

The paper is organized as follows. Section 2 discusses related works. Section 3 formulates the problem and discusses the audio-visual trajectory estimation, whereas Section 4 presents the proposed interaction event recognition approach. Experimental results are discussed in Sec. 5. Finally, in Sec. 6 we draw conclusions.

2. RELATED WORK

Interaction Event Recognition (IER) can be modeled as a random process that is segmental in nature. Various spatio-temporal features are extracted from video sequence to identify these interactions [11]. Decisions from single actor action classifiers [4] trained on these features can be combined to perform interaction recognition by formulating a decision profile matrix (*group member* $\times$ *action performed*) [13]. The product, sum, min and max rule applied in the column space of the decision profile then gives insight about interactions. However, this method is only applicable to symmetric interactions such as for example *hand shaking* and *hugging*. For asymmetric interactions, the decision profile is divided into symmetric and asymmetric block matrices [13], which provide a redundant representation in case of symmetric interactions. A multi-channel Support Vector Machine (SVM) is instead trained on spatio-temporal features to localize actions and interactions simultaneously [4]. This method also uses movie scripts to automatically generate training data for SVM and does not scale to complex scenarios. Although various types of Hidden Markov Models (HMM) have been used for event recognition [1, 3, 6, 8, 5], standard HMMs can only model a state sequence of a single process. Because objects move simultaneously based on their intentions, their actions depend also upon the behaviours of other objects. For this reason, monitoring a single target separately may not provide the complete information about its state. Variants of HMM such as

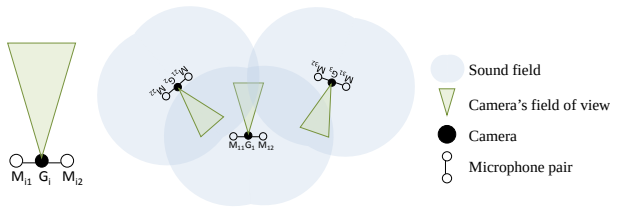

Fig. 1. Stereo Audio Cycloptic Vision (STAC) sensors and their coverage area (Key. M_{i1} and M_{i2}: microphone pair; G_i: camera).

This work was supported in part by the EU, under the FP7 project APIDIS (ICT-216023).

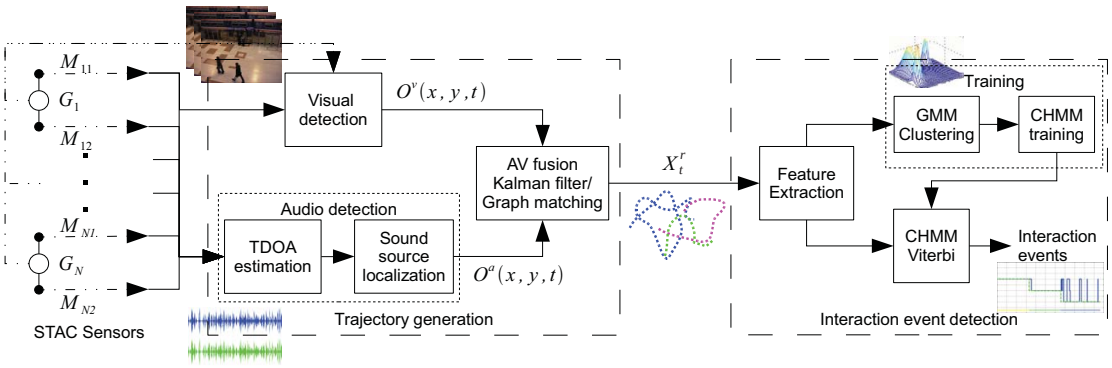

Fig. 2. Block diagram of the proposed framework for interaction event recognition in wide areas (Key. STAC: Stereo Audio Cycloptic Vision; TDOA: time difference of arrival; M_{i1} and M_{i2}: microphone pair; G_i: camera).

Multi-Observation-Mixture+Counter HMM allow representation of multiple observations of different objects, for example based on object silhouette as feature [1]. However, object silhouettes are only available in the cameras' FOV and are difficult to segment correctly. The dependency on silhouettes can be avoided using trajectories. Since tracks of different targets are likely to have varying lengths, then Variable-length HMMs (VLHMM) can be used for example in modeling interactions among cars on highways [3]. However, VLHMMs are difficult to train as the structure of the model is unknown. Finally, Coupled Hidden Markov Models (CHMMs) have received significant attention [6, 8, 5] to model group activity, as they allow modeling the full coupling between the processes and can be solved in polynomial time using dynamic programming. Events like *walk*, *approach* and *chat* have been detected using CHMM [6, 5] with synthetic data. CHMMs have also been applied to medical data to detect changes in heart beats during sleep [8] and for gesture recognition [11].

3. AUDIO-VISUAL TRAJECTORY ESTIMATION

Let a wide area be monitored by a set $G = \{G_1, \ldots, G_N\}$ of N cameras with non-overlapping FOVs. Let each camera be equipped with a microphone pair, with $M = \{M_1, \ldots, M_N\}$ being the set of N microphone pairs, where $M_i = (M_{i1}, M_{i2})$. We assume that the microphones' sound field is wider than the corresponding cameras' FOV and that the sound field of multiple microphone pairs M_i overlap each other (Fig. 1). Our goal is to model the interaction between targets O^p and O^q in regions within *as well as* outside the cameras' FOV. To achieve this goal, the trajectory X_t^r up to time t representing the movement of each object $r \in \{p, q\}$ needs to be estimated over the entire audio-visual sensor network through *audio-audio* and *audio-visual* fusion. The problem of trajectory estimation in audio-visual sensor networks can be divided into (i) visual trajectory estimation in the cameras' FOV; (ii) audio-audio fusion in the unobserved regions for target localization and (iii) audio-visual fusion in regions observed by both audio and visual sensors.

We estimate trajectories in regions observed by the cameras using a multi-target tracker based on graph matching [12], which is robust to short-term occlusions. We perform *audio-audio* fusion from multiple microphone pairs by fusing localization information from multiple pairs. The localization information from each microphone pair is estimated by computing the direction of arrival θ of the sound followed by estimating the intersection of the arrival angle lines [9] between two microphone pairs (Fig. 3). Note that this localization can be erroneous and the error increases as the angle between the two intersecting lines $\rho \to 0$ or $\rho \to 180$. The minimum localization error is achieved at $\rho = 90$. For this reason, we ignore the estimation of a microphone pair for $147^o < \rho$ or $\rho < 33^o$ and the information from the other microphone pairs is instead used. Audio performance also decreases as the target moves closer than 5m from the sensor, as the assumption of parallel sound waves in time difference of arrival (TDOA) estimation will no longer be valid. In this case, as the microphone pair is unable to provide the localization information, we estimate the trajectory using a first-order motion model. Finally, we perform the *audio-visual fusion* within Kalman filtering with a weighted sum of the two measurements [10] that only penalizes audio detections in overlapping regions and gives a weight no smaller than 0.5 to the video modality, when available.

The resulting trajectories X_t^r, $r \in \{p, q\}$ of any two targets O^p and O^q provide the estimated positions (x^p, y^p) and (x^q, y^q) at each time t. These trajectories *within* and *outside* the cameras' FOV are then used for interaction recognition, as discussed in the next section.

4. INTERACTION EVENT RECOGNITION

Let us consider the following interactions of interest: *follow-reach-go together* (E1), *approach-meet-go separately* (E2), *approach-meet-go together* (E3), *change direction-approach-meet-go separately* (E4) and *change direction-approach-meet-go together* (E5). These interactions are of interest for annotating behaviours in a scene [6] and are composed by combinations of the following atomic events: *approaching*, *meeting*, *going together*, *going separately* and *following*. To recognize the interactions Ei using the trajectories estimated as described in Section 2, we analyze a set of features of target pairs, p and q, represented as a feature vector $\mathbf{f}_t^{(p,q)}$. We consider the relative direction, the relative distance and its derivative as well as the magnitude of velocity of each object. The relative direction, $\phi^{(p,q)}$, between two targets is defined as

$$\phi_t^{(p,q)} = tan^{-1}\left(\frac{y_t^p - y_{t-1}^p}{x_t^p - x_{t-1}^p}\right) - tan^{-1}\left(\frac{y_t^q - y_{t-1}^q}{x_t^q - x_{t-1}^q}\right). \quad (1)$$

To disambiguate between *going together* and *going separately*, we use the relative distance $d^{(p,q)}$ and its derivative $\dot{d}_t^{(p,q)}$:

$$d_t^{(p,q)} = \sqrt{(x_t^p - x_t^q)^2 + (y_t^p - y_t^q)^2}. \quad (2)$$

To differentiate between *following* and *approaching*, we use the relative direction and distance based on the current position of two in-

teracting targets and not from a fixed reference point. We also use individual object features in the form of magnitude of velocity $\nu_t^{(p)}$:

$$\nu_t^{(p)} = \sqrt{\nu_{x,t}^{(p)^2} + \nu_{y,t}^{(p)^2}}, \tag{3}$$

where $\nu_{x,t}^{(p)}$ and $\nu_{y,t}^{(p)}$ are the horizontal and vertical components of the velocity. Similarly for $\nu_t^{(q)}$.

The overall feature vector for any two targets O^p and O^q is therefore

$$\mathbf{f}_t^{(p,q)} = \left(\nu_t^{(p)},\ \nu_t^{(q)},\ \phi_t^{(p,q)},\ d_t^{(p,q)},\ \dot{d}_t^{(p,q)}\right). \tag{4}$$

Now that the features are defined, let there exist a full coupling between the target states: then the interactions between O^p and O^q is modeled as 1-lag, 2-chain Coupled Hidden Markov model (CHMM). Let in CHMM O^p and O^q represent the set of observations; S^p, S^q be the set of states for the two chains and $P(s_{t+1}^i|s_t^j)$, $i, j \in \{p, q\}$ be the transition probabilities among these states. We train a CHMM based on the Expectation Maximization (EM) algorithm initialized using GMM clustering to avoid local maxima by assigning means (μ_{GMM}) and covariance Σ_{GMM} of the mixture model to CHMM states [8]. Because it is difficult to obtain enough training samples from real dataset, the training data for the interactions of interest is generated synthetically using a first-order motion model so that the computed trajectories not only have the desired interactions, but also resembles target motion in many real-world scenarios. This allows us to use the CHMM model trained on synthetic data on real trajectories, without retraining. Real tracks are normalized by projecting them into the environment where the synthetic data is generated (i.e. a 4×4 units area).

The clustering of features is performed on a 5-dimensional feature space defined by $\mathcal{F} : X_t^i \times X_t^j \to \Re^5,\ \forall\, i, j$. The CHMMs are trained on these features using the maximum a posterior (MAP) approach, instead of Maximum Likelihood (ML), as MAP incorporates the prior distribution over the quantities to be estimated to achieve a better optimization. Using the EM algorithm MAP approach [8], we aim to maximize

$$\mathcal{Q}(\lambda) = \int Q(S|\lambda) \log P(S, O|\lambda) dS + P(\lambda), \tag{5}$$

where $Q(S|\lambda)$ is the probability of a state given the model λ; λ consists of the state transitions, the initial state probabilities and the emission probabilities; and $P(\lambda)$ is the probability of the model parameters.

We perform interaction recognition by projecting the trajectory onto $\mathcal{F}$ and then applying the CHMM Viterbi decoding using the trained model parameters. The decoding strategy is preferred here over the evaluation strategy because of its flexibility, as the former does not require event templates to be recognized and allows generating a sequence of activities performed by the targets, as evaluated in the next section.

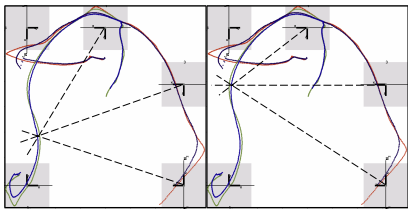

Fig. 3. Localization using arrival angle estimation from multiple STAC sensors (Key. Red and green lines: ground truth; blue and the black lines: estimated trajectories; grey squares: cameras' FOV; black dashed lines: audio source localization result).

Table 1. Accuracy comparison for interaction recognition using the four approaches under analysis on the five events of interest

		E1	E2	E3	E4	E5
DBN	μ	.8069	.7084	.7706	.4702	.4403
	σ	.1300	.1416	.1045	.1248	.1481
HMM	μ	.8585	.7791	.6967	.7987	.7555
	σ	.0717	.0743	.2210	.0546	.0808
CHMM-ML	μ	.8563	.7881	.5419	.7976	.7527
	σ	.0798	.0717	.1861	.0574	.0851
CHMM-MAP	μ	.8665	.8688	.7650	.8376	.8049
	σ	.0682	.0575	.1175	.0557	.0804

5. EXPERIMENTAL RESULTS

We compare the proposed IER approach with three alternative algorithms based on a Dynamic Bayesian Network (DBN), an HMM and a CHMM Maximum Likelihood (CHMM-ML) approach, all on the same feature space used by the proposed approach, CHMM-MAP. In the DBN the states are defined as a set of discrete and continuous random variables and the transition and observation models are defined as a product of the conditional probability distributions. The methods are evaluated on a synthetic dataset (D1) consisting of 100 trajectories pairs for each event and of approximately 750 data points per trajectory and on a real dataset from soccer matches (D2 and D3). D2 consists of 277 frames (25 fps, resolution 1440×537 pixels) while D3[1] consists of 6002 frames (25 fps, resolution 1920×1080 pixels). The synthetic data are equally divided into training and test sets containing interactions E1 to E5. The real dataset inherently contains the interactions that are modeled by our approach. An audio signal was added to selected trajectories from the real data, followed by the re-estimation of trajectories using audio-visual data. The accuracy of the recognition results is computed as

$$\wp = \frac{\zeta(GT_t \cap AR_t)}{N_{GT}}, \tag{6}$$

where AR_t is the automatic recognition result, N_{GT} is the duration of the event in the ground truth, ζ gives the number of elements in the interaction set and t_0 is the starting time of the event. GT_t is the ground truth and it is a discrete variable containing the state of the interacting targets at each time instance during the event span $t = t_0, \cdots, t_0 + N_{GT}$.

Table 1 shows the comparison of the IER via state estimation using Viterbi decoding with the three alternative methods. CHMM-MAP trained on the selected feature set achieves the highest mean accuracy μ. On the one hand, DBN has a very low accuracy for E4 and E5 as these interactions present intervals of random walk at the beginning and at the end. On the other hand, DBN has the highest accuracy for E3 as it has the smallest amount of random walk. The standard deviation σ of CHMM-MAP is the lowest among all approaches except for E3 and E4. For E3, DBN has a lower standard deviation, whereas for E4 the HMM has a lower σ. For E1, E2, E4 and E5 both CHMMs have the highest accuracy. In fact, the assumption of local interaction in CHMMs (the hidden variables are assumed to interact locally with their neighbors) is consistent with the definition of interactions between people thus resulting in improved performance compared to DBN, which is not based on such assumption.

[1]Raw videos courtesy of Institute of Intelligent Systems for Automation - C.N.R., Bari, IT. *http://www.issia.cnr.it*

Figure 4(a,b) shows E2 where two players from opposite teams *approach* the ball almost simultaneously, *stay together* while tackling the ball. Then the ball is kicked away and the two players *go separately*. The detection and tracking results are shown in Fig. 4(a) where the 22 players are detected and tracked. The two players interacting with each other while trying to get the ball are shown in a magnified section on the corner of the image. The ground truth and the resulting state sequence are shown in Fig. 4(b). The horizontal color bar indicates state 2 (*approaching*) in light green, state 3 (*meeting*) in brown and state 1 (*going separately*) in blue. The state sequence generated by CHMM-MAP is shown with dotted and dashed lines which coincide with the ground truth by 97.45%. The few flickerings during state 3 are due to the fact that targets never stopped completely and the state of *staying* is not a stationary state. E1 is shown in Fig. 4(c,e). The player from team 1 (navy blue) *follows* the player of team 2 (white) who is following the ball. The player of team 1 increases speed and gets closer to the player of team 2 and then they *go together* towards the ball. This interaction sequence detected with 97.95% similarity with ground truth (Fig. 4(d,f)). The *meeting* state is detected earlier due to blob merging and is a very short state as the targets hardly stop and continue chasing the ball.

6. CONCLUSIONS

We have presented an end-to-end event recognition algorithm for multimodal sensor networks. The extracted trajectory estimated by fusing information from multiple modalities gives a complete trajectory estimation of the target in a wide environment and helps in better understanding interactions. Our approach detects interaction events using a CHMM-MAP trained on a feature set that encapsulates the dynamics of interacting objects in real scenarios characterized by high variability in targets' directions. The performance of the proposed algorithm was demonstrated on both synthetic and real data, without the need of additional training. Future work includes the extension of the proposed approach to model interactions involving groups and its use for sensor selection.

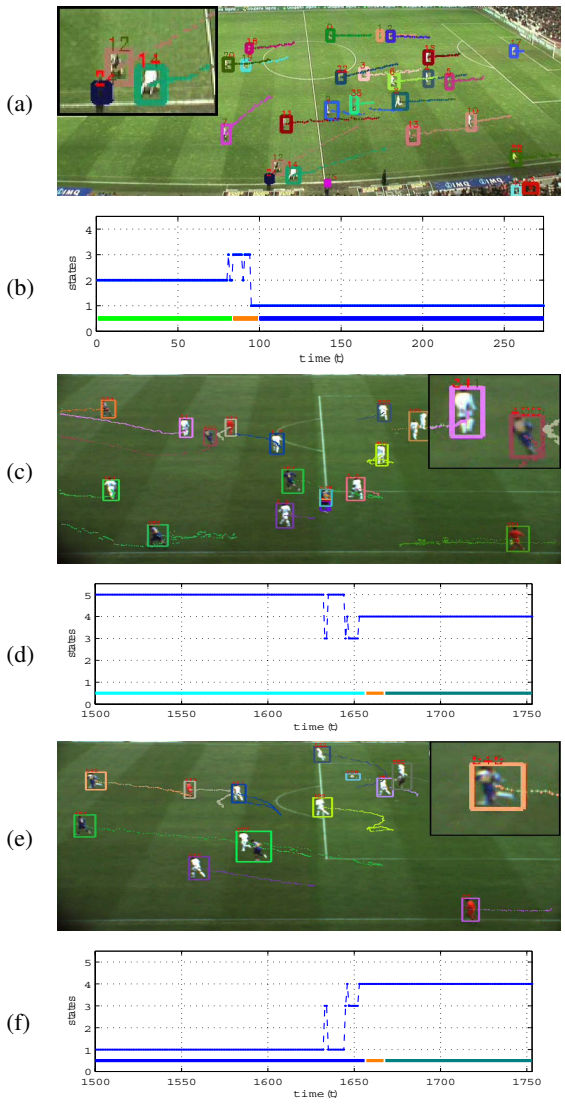

Fig. 4. Interaction event recognition results on a real scenario. (a) Frame 50 of soccer match with *approach-meet-go separately* interaction (E2) among targets shown on the magnified area. (b) CHMM-MAP generated sequence of interaction and the ground truth. (c,e) Frames 1640 and 1720 of a soccer match with showing *follow-reach-go together* interaction (E1). (d,f) CHMM-MAP generated sequence of interactions with the ground truth. (Key for the plots: light green: *approaching*; brown: *meeting* and *waiting to meet*; blue: *walking/going separately* cyan: *follow*; dark green: *going together*; dark blue: *going separately*).

7. REFERENCES

[1] M. Brand and V. Kettnaker. Discovery and segmentation of activities in video. *IEEE Trans. on PAMI*, 22:844–851, Aug 2000.

[2] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE Trans. on PAMI*, 30(2):267–282, Feb 2008.

[3] A. Galata, A. Cohn, D. Magee, and D. Hogg. Modeling interaction using learnt qualitative spatio-temporal relations and variable length Markov models. In *ECAI*, Lyon, FR, Jul 2002.

[4] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE CVPR*, Anchorage, AK, USA, Jun 2008.

[5] P. Natarajan and R. Nevatia. Coupled hidden semi Markov models for activity recognition. In *IEEE Int. WMVC*, Austin, TX, USA, Feb 2007.

[6] N. Oliver, B. Rosario, and A. Pentland. A Bayesian computer vision system for modeling human interactions. *IEEE Trans. on PAMI*, 22:831–843, Aug 2000.

[7] A. Rahimi, B. Dunagan, and T. Darrell. Simultaneous calibration and tracking with a network of non-overlapping sensors. In *IEEE CVPR*, volume 1, Washington, DC, USA, Jun 2004.

[8] I. Rezek, M. Gibbs, and S. J. Roberts. Maximum a posteriori estimation of Coupled Hidden Markov models. *Journal of VLSI Sig. Proc. Sys.*, 32(1-2):55–66, 2002.

[9] R. Schmidt. A new approach to geometry of range difference location. *IEEE Trans. on Aero. Elec. Sys.*, AES-8(6):821 –835, Nov 1972.

[10] M. Taj and A. Cavallaro. Audio-assisted trajectory estimation in non-overlapping multi-camera networks. In *IEEE ICASSP*, Taipei, 2009.

[11] M. Taj and A. Cavallaro. *Intelligent Multimedia Analysis for Security Applications*, chapter 2: Recognizing Interactions in Video, pages 29–57. Springer Verlag GmbH, 2010.

[12] M. Taj, E. Maggio, and A. Cavallaro. Multi-feature graph-based object tracking. In *CLEAR, LNCS 4122*, Southampton, UK, Apr 2006.

[13] D. Waltisberg, A. Yao, J. Gall, and L. V. Gool. Variations of a hough-voting action recognition system. In *IEEE ICPR*, Istanbul, 2010.