Audio–Visual Synchronisation for Speaker Diarisation

Giulia Garau, Alfred Dielmann and Hervé Bourlard

Idiap Research Institute - CP592, 1920 Martigny, Switzerland
Ecole Polytechnique Federale de Lausanne - 1015 Lausanne, Switzerland
giuliagarau@yahoo.com, a.dielmann@gmail.com, herve.bourlard@idiap.ch

Abstract

The role of audio–visual speech synchrony for speaker diarisation is investigated on the multiparty meeting domain. We measured both mutual information and canonical correlation on different sets of audio and video features. As acoustic features we considered energy and MFCCs. As visual features we experimented both with motion intensity features, computed on the whole image, and Kanade Lucas Tomasi motion estimation. Thanks to KLT we decomposed the motion in its horizontal and vertical components. The vertical component was found to be more reliable for speech synchrony estimation. The mutual information between acoustic energy and KLT vertical motion of skin pixels, not only resulted in a 20% relative improvement over a MFCC only diarisation system, but also outperformed visual features such as motion intensities and head poses.

Index Terms: multimodal speaker diarisation, audio–visual speech synchrony, multiparty meetings, mutual information, canonical correlation analysis

1. Introduction

In this paper we investigate the speaker diarisation task, using both audio and audio–visual synchrony cues. The goal of speaker diarisation is estimating "who spoke when" [1]. A robust speaker diarisation approach is beneficial for applications such as: automatic speech recognition, dominance detection, automatic role recognition, and addressee identification. Most speaker diarisation systems work in two steps: the audio stream is classified into speech and non-speech segments (speech/non–speech detection), then, speech segments uttered by the same speaker are grouped (clustering) [2]. Speaker clustering has been traditionally approached employing only acoustic cues [2, 1]. However there is recently an increasing number of works which investigate the role of visual cues (such as motion and eye gaze) for speaker diarisation [3, 4, 5, 6].

Audio–visual synchrony features are motivated by the importance of audio–visual co–occurrences for sound localisation: facial movements are strongly correlated with speech acoustics which in turn is also correlated with vocal tract movements [7]. Hershey and Movellan [8] proposed to use an audio–visual synchrony measure for speaker localisation, based on the Mutual Information (MI) of acoustic energy and individual pixel luminance variations (considering the whole image). This work was further extended by Nock et al. [9] employing visual features based on lip detection. Slaney and Covell [10] adopted Canonical Correlation Analysis (CCA) to maximise the audio–visual correlation and exploited this maximised quantity as a measure of audio–visual synchronisation. However, the data used in these studies is generally small and constrained to non–overlapping speech and frontal face views. Mutual information, estimated both on the facial region and on the whole video–frame, was adopted by Noulas et al. [3] for multimodal speaker diarisation of a single meeting recording.

In this paper two audio–visual synchrony measures are compared: Mutual Information and Canonical Correlation Analysis. Moreover we investigate the synchrony of different acoustic and visual features: acoustic energy and Mel Frequency Cepstral Coefficients (MFCCs) for the acoustic domain; individual pixel luminance variations and motion tracking for the visual domain. In particular we propose the adoption of the Kanade Lucas Tomasi (KLT) [11] tracking algorithm and skin colour modelling to estimate facial motion during the computation of audio–visual synchrony. Since motion is measured only in a selected set of points, the estimation of MI/CCA on KLT tracked feature points is computationally advantageous compared to the computation of synchrony considering the whole image [3, 8]. Thanks to the KLT we are able to decompose the movement into its horizontal and vertical components. We found that, being speech production more correlated with vertical movements of lips and chin, the vertical component is a better cue for the computation of audio–visual synchrony. Moreover our novel KLT based technique results in consistent speaker diarisation improvements: MI features based on the KLT vertical motion component outperform the MI estimated considering pixel luminance variations over time.

The audio–visual synchrony features are integrated during the clustering phase of the speaker diarisation. We test our novel approach on more than 5 hours of unconstrained multiparty meetings: an interesting and challenging domain, both from the acoustic and visual point of view. Meeting participants have variable length speaker turns, their voices sometimes overlap, and they can move freely in the room (for example to go to the whiteboard). One of the goals of this paper is to compare audio–visual synchrony features with the visual cues we investigated in [6]: motion and head pose features. The use of head poses, which can be seen as an approximation of eye–gaze, is motivated by language and social psychology studies on the role of gaze in a conversation [12]: listeners are likely to look at the person who is talking and they request turn shifts using gaze; speakers are likely to look at their addressee and to shift their attention towards the next speaker before a speaker turn occurs. Motion intensity features take into account speaker movements for speech production and gestures [12]. In our speaker diarisation experiments we found that audio–visual synchrony features are more robust than motion and head pose features.

This paper is structured as follows: in Section 2 the adopted speaker diarisation framework is outlined; Section 3 describes the data; Section 4 and 5 outline the audio and visual features respectively. In particular the proposed audio–visual synchrony features are introduced in Section 6. Section 7 discusses speaker diarisation experiments and results, finally Section 8 summarises this work highlighting the most important findings.

2. Speaker diarisation engine

The work presented in this paper is based on the ICSI speaker diarisation system [2]. This system uses the following bottom-up agglomerative clustering approach. Speaker clusters are modelled with an ergodic Hidden Markov Model (HMM). Each state (corresponding to a single speaker cluster) is modelled as a sequence of hidden substates sharing the same Gaussian Mixture Model (GMM). In order to enforce a minimum duration constraint of 2.5 seconds the same substate is duplicated several times. The first step of the ICSI speaker diarisation system is the Speech/Non-Speech detection [2]; then, processing only the speech frames, K initial clusters are created uniformly partitioning the speech frames in $K = 16$ clusters of equal length. A GMM is trained for each initial speaker cluster. Three processing steps are then iterated: Viterbi decoding using the current ergodic HMM, training of a new GMM for each speaker cluster using the newly estimated segmentation, and cluster merging. For each iteration, the most similar cluster pair is found according to a score based on the Bayesian Information Criterion (BIC). This is obtained measuring the difference between the log likelihood of the model trained jointly on the data belonging to the two clusters (θ) and the sum of the log likelihoods of the models of the two clusters (θ_a and θ_b) modelled independently. It is also assumed that the complexity of the model θ is equal to the sum of the complexities of the models θ_a and θ_b.

The integration of multiple feature streams (e.g. two streams) is performed by training separate GMMs for each stream. The two streams are combined both during Viterbi segmentation and clustering, computing the total log likelihood as a weighted sum of the likelihood of the two separate models. In our experiments the first stream is always represented by 19 MFCCs and, being this the most informative modality for speaker diarisation, a weight of 0.9 was assigned to MFCCs while 0.1 was adopted for the additional feature stream.

3. Data

Experiments were performed on a subset of the AMI meeting corpus[1] [13]. This multimodal collection of four participant meetings was recorded in rooms instrumented with a set of synchronised devices, as shown in Figure 1. We used the 8-element circular table-top microphone array for audio feature extraction, the two side-cameras to extract head poses, and the four individual closeup cameras to extract motion activity features and KLT features (Section 5.1 and 5.3 respectively). We selected the 11 meetings with the richest annotation [6]. These meetings offer a variety of challenges both from the audio and the video point of view (overlapping speech, moving speakers, and poor head resolution). We can distinguish between static meetings, where people seat during the entire meeting, and dynamic meetings, where people leave their seats to go to the whiteboard or the slide–screen.

4. Audio features

Beamforming was adopted to reduce the $d = 8$ microphone array signals (Section 3) to a single channel with enhanced sensitivity in the direction of the desired signal. To perform this task we used the *Beamformit* tool[2] [14], based on the delay and sum algorithm. First a reference channel is chosen so that the average cross-correlation with the other channels is maximised.

[1] Publicly available from `http://corpus.amiproject.org`

[2] `www.icsi.berkeley.edu/~xanguera/beamformit/`.

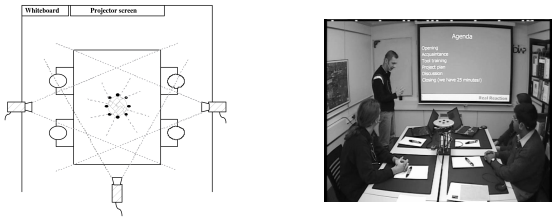

Figure 1: Meeting room setup.

With respect to this reference channel, $(d - 1)$ Time Delays of Arrival are estimated using the generalised cross–correlation with phase transform method. TDoAs are then used for delay and sum beamforming. On the beamformed signal we compute 19 MFCCs as acoustic features for the speaker clustering (Section 2). We also extract two separate sets of acoustic features $f_a(t)$ for the computation of audio–visual synchrony measures (Section 6): the acoustic energy and its combination with the MFCCs C1, C2 and C3.

5. Visual Features

5.1. Motion Intensity features

Motion intensity features were extracted on each of the four closeup videos as the average pixel by pixel luminance difference of subsequent frames [6]. These motion features were used in the baseline speaker diarization experiments presented in Section 7. The luminance variations between adjacent video frames, for each skin color like pixel in the closeup camera recording, were employed as visual features $f_v(x, y, t, k)$ for the audio–visual synchrony estimation (Section 6). Skin color detection is performed using the YUV histogram model trained on Banca data, provided by the Torch3vision toolkit[3].

5.2. Head pose likelihood features

We employed the head pose likelihoods estimated during the VFoA detection [15]. They represent the probability $P(O(i,t)|S_k)$ of the observed head pose $O(i,t)$ of meeting participant i given that his visual focus S_k is participant k at time t. While listening, people are more likely to look at the person which is speaking, thus head pose features were defined as the sum of the probabilities that each meeting participant i is looking at the meeting participant k [6]: $f_{headpose}(k,t) = \sum_{i=1\, i\neq k}^{i=I} \frac{P(O(i,t)|S_k)}{I-1}$ where I is the number of meeting participants.

5.3. Kanade Lukas Tomasi based features

The motion intensities described in the Section 5.1 measure the luminance variation of each skin–like pixel in the image. In order to assess the synchrony of facial movements with speech acoustics, we also experimented with motion estimates based on a widely adopted tracking technique: the Kanade Lucas Tomasi (KLT) algorithm [11]. KLT works in two steps: features (i.e. points of interest) are initially selected in the image and then tracked. Given a feature point with luminance $I(x, y, t)$ in the image frame t, feature tracking goal is to estimate its correspondent position $(x + dx, y + dy)$ in the following frame $t + \tau$, assuming that $I(x, y, t) \approx I(x + dx, y + dy, t + \tau)$. This optical flow estimation problem can be solved using the Kanade Lucas

[3] `torch3vision.idiap.ch`.

Tomasi method. KLT assumes the displacement components in the horizontal and vertical directions (d_x, d_y) constant in a neiborhood $W : (x \pm w_x, y \pm w_y)$, and minimises the equation:

$$\epsilon(d_x, d_y, t) = \sum_{(x,y)\in W} (I(x, y, t) - I(x + d_x, y + d_y, t + \tau)).$$

KLT selects feature points which can be tracked well [11], usually consisting in corners and salt–and–pepper textures. In our experiments we used the S. Birchfield implementation of KLT[4]. Since we are interested in skin–like areas, we also used skin color detection (Section 5.1) to track only those feature points which are likely to be skin. Given the KLT tracked skin–like feature points (x, y) in each closeup camera k at time t, we consider the displacement $d_x(x, y, t, k)$ and $d_y(x, y, t, k)$ over time as visual features $f_v(x, y, t, k)$.

6. Mutual Information and Canonical Correlation Analysis measures

Considering the image points (x, y)[5] of the closeup camera k and temporal frame t, we define two measures of audio–visual synchrony $\phi(f_a(t), f_v(x, y, t, k))$ based on: mutual information (MI) and canonical correlation analysis (CCA).

Mutual information is estimated according to the definition of Hershey and Movellan [8]. Let $f_a(t)$ be the audio features of dimension N_a and $f_v(x, y, t, k)$ be the video features of dimension N_v. Assuming $f_a(t)$ and $f_v(x, y, t, k)$ independently and jointly Gaussian in a window $[t - \Delta t/2, t + \Delta t/2]$ (in our experiments $\Delta t = 0.5\ secs$), the mutual information can be computed as:

$$\phi_{MI}(f_a(t), f_v(x, y, t, k)) = \frac{1}{2} log \frac{|\boldsymbol{\Sigma_{aa}(t)}|\,|\boldsymbol{\Sigma_{vv}(x, y, t, k)}|}{|\boldsymbol{\Sigma_{av}(x, y, t, k)}|}.$$

$\boldsymbol{\Sigma_{aa}(t)}$ and $\boldsymbol{\Sigma_{vv}(x, y, t, k)}$ are respectively the covariance matrices for audio and visual features, and $|\bullet|$ is the matrix determinant. The joint audio–visual covariance matrix is:

$$\boldsymbol{\Sigma_{av}(x, y, t, k)} = \begin{bmatrix} \boldsymbol{\Sigma_{aa}(t)} & \boldsymbol{C_{av}(x, y, t, k)} \\ \boldsymbol{C_{va}(x, y, t, k)} & \boldsymbol{\Sigma_{vv}(x, y, t, k)} \end{bmatrix},$$

where $\boldsymbol{C_{av}(x, y, t, k)} = \boldsymbol{C_{va}(x, y, t, k)}^T$ is the between-sets covariance matrix.

Canonical correlation analysis [16] aims at finding the projections $\boldsymbol{W_a}$ and $\boldsymbol{W_v}$ such that the correlation ρ between the features $f_a(t)$ and $f_v(x, y, t, k)$ is maximised. Canonical correlations can be estimated (using the same window $\Delta t = 0.5\ secs$ adopted for the MI estimation) by solving the following eigenvalue equation:

$$\boldsymbol{\Sigma_{aa}}^{-1} \cdot \boldsymbol{C_{av}} \cdot \boldsymbol{\Sigma_{vv}}^{-1} \cdot \boldsymbol{C_{va}} \cdot \boldsymbol{W_a} = \boldsymbol{\rho}^2 \cdot \boldsymbol{W_a}$$

where the squared canonical correlations $\boldsymbol{\rho}^2$ and $\boldsymbol{W_a}$ can be computed as the eigenvalues and eigenvectors of $\boldsymbol{\Sigma_{aa}}^{-1} \cdot \boldsymbol{C_{av}} \cdot \boldsymbol{\Sigma_{vv}}^{-1} \cdot \boldsymbol{C_{va}}$. We define the audio–visual correlation measure ϕ_{CCA} as the sum of the eigenvalues ${\rho_i}^2(f_a(t), f_v(x, y, t, k))$ [17]:

$$\phi_{CCA}(f_a(t), f_v(x, y, t, k)) = \sum_{i=1}^{N} {\rho_i}^2(f_a(t), f_v(x, y, t, k))$$

[4] www.ces.clemson.edu/~stb/klt.

[5] The (x, y) image points, on which the audio–visual synchrony is estimated, include: all the skin pixels in the motion intensity experiments; all the skin–like tracked feature points when KLT features are adopted.

Table 1: MI/CCA synchrony measures ϕ obtained from 10 different combinations of acoustic f_a and visual f_v features.

$\phi(f_a(t), f_v(x, y, t, k))$	$f_a(t)$	$f_v(x, y, t, k)$
MI-A CCA-A	acoustic energy	motion intensity (skin pixels only)
MI-Bx CCA-Bx	acoustic energy	KLT tracking horizontal displacement d_x
MI-By CCA-By	acoustic energy	KLT tracking vertical displacement d_y
MI-Cx CCA-Cx	ac.energy C1,C2,C3	KLT tracking horizontal displacement d_x
MI-Cy CCA-Cy	ac.energy C1,C2,C3	KLT tracking vertical displacement d_y

where N is the minimum between N_a and N_v.

Synchrony measures: were estimated for each closeup camera k and for each time frame t. Given the set $M_{t,k} = \{\forall(x, y) : \phi(f_a(t), f_v(x, y, t, k)) > 0\}$ of points (x, y), we define the synchrony measure $synchrony(t, k)$ as the average of the ϕ_{MI} or the ϕ_{CCA} measures over this set:

$$synchrony(t, k) = \frac{1}{\#M_{t,k}} \sum_{(x,y)\in M_{t,k}} \phi\left(f_a(t), f_v(x, y, t, k)\right),$$

where $\#M_{t,k}$ is the cardinality of the set $M_{t,k}$.

The resulting synchrony features measure to what extent the motion features $f_v(x, y, t, k)$ could be predicted given the acoustic features $f_a(t)$, and viceversa.

Audio–visual features: we have estimated ϕ_{MI} and ϕ_{CCA} using different audio features $f_a(t)$ (Section 4) and video features $f_v(x, y, t, k)$ (Section 5.1 and 5.3) as outlined in table 1.

7. Experimental Results

Speaker diarisation performances, in terms of Diarisation Error Rate (DER), were evaluated using the tools provided by NIST [6]. DER is defined as the sum of the Speech/Non-Speech error and the speaker error percentage. An average Speech/Non-Speech detection error of 13.9% is shared across all the experimental setups presented in this paper; thus we can only aim at reducing the speaker error percentage.

In section 6 we outlined two audio–visual synchrony measures: Mutual Information (MI) and Canonical Correlation Analysis (CCA). Each of these measures can be estimated using different combinations of acoustic features $f_a(t)$ and visual features $f_v(x, y, t, k)$, such as: acoustic energy, 3 MFCCs (C1, C2, C3), pixel luminance variation (i.e. motion intensity), horizontal and vertical displacements d_x and d_y from the KLT tracking (Section 5.3). Table 1 outlines the feature combinations employed by the speaker diarisation experiments of table 2. The multistream speaker diarisation system jointly modelled synchrony and MFCC features, associating them to two independent feature streams (Section 2). For comparison, on top of table 2, we also report baseline results [6] using a single MFCC stream, and its combination with motion intensity and head pose features (Section 5.1 and 5.2 respectively). We report results for the whole set and in brackets for static and dynamic meetings.

Mutual information features **MI-A**, estimated using acoustic energy and motion intensities of the skin pixels , resulted in 27.8% of DER. Therefore **MI-A** outperforms the motion intensities baseline system (28.6%), in particular on dynamic

[6] www.nist.gov/speech/tests/rt/2006-spring/.

Table 2: DER results for the whole dataset (Total), and in square brackets for static and dynamic meetings (Section 3).

1^{st} stream	2^{nd} stream	Total [Static Dynamic]
MFCC	—	31.0 [14.7 36.5]
MFCC	Motion intensities	28.6 [13.0 33.8]
MFCC	Head pose features	26.6 [13.1 31.0]
MFCC	MI-A	27.8 [13.5 32.6]
MFCC	CCA-A	29.7 [22.9 32.7]
MFCC	MI-Bx	29.3 [13.4 34.5]
MFCC	MI-By	**24.7 [13.5 28.5]**
MFCC	CCA-Bx	27.5 [12.9 32.4]
MFCC	CCA-By	26.5 [13.1 31.0]
MFCC	MI-Cx	29.9 [13.3 35.5]
MFCC	MI-Cy	26.0 [12.8 30.4]
MFCC	CCA-Cx	30.3 [13.4 35.9]
MFCC	CCA-Cy	27.4 [13.4 32.1]

meetings (Section 3). Mutual information features **MI-By**, estimated using acoustic energy and KLT vertical displacements, provided the best speaker diarisation result (24.7% DER). In this setup the Mutual Information is computed considering only the skin pixels tracked by KLT; besides attaining the best absolute DER, this approach is also more computationally efficient than **MI-A**. Mutual information synchronies **MI-Cx** and **MI-Cy** are estimated employing the acoustic energy and the first three MFCCs as $f_a(t)$, and the KLT displacements d_x and d_y as $f_v(x, y, t, k)$. In particular **MI-Cy** attained good diarisation performances (26.0%) outperforming both **MI-Bx** and **MI-Cx**.

Analogue experiments were also performed using the CCA synchrony measure (instead of MI). In this case the best result **CCA-By** (26.5%) is achieved when the acoustic energy is used as $f_a(t)$ together with the KLT vertical displacement d_y as $f_v(x, y, t, k)$. Even if the CCA synchrony measures did not outperform their MI counterparts, they provided better results than using motion intensities.

The use of the displacement along the vertical axis (**MI-By**,**MI-Cy**,**CCA-By**,**CCA-Cy**) provided better diarisation performances compared to the horizontal displacement (**MI-Bx**,**MI-Cx**,**CCA-Bx**,**CCA-Cx**) in all the cases, both using MI and CCA synchrony measures. Facial movements along the vertical axis are more correlated with speech production (e.g. lips and chin movements). Considering only this component we were able to neglect horizontal movements (such as head shaking and horizontal head rotations) which may not be relevant for audio–visual speech synchrony.

8. Conclusions

The aim of this paper is the investigation of two audio–visual synchrony measures, mutual information and canonical correlation analysis, in the context of speaker diarisation of unconstrained multiparty meetings. These measures are motivated by the correlation between facial movements and speech acoustics. We compared the estimation of audio–visual synchrony using different feature sets: energy and MFCCs as acoustic features; motion intensities (based on luminance differences) and motion features based on KLT and skin detection as visual cues. In particular KLT features allowed to isolate the horizontal and the vertical motion components. We found that the vertical movement, being more correlated with speech production, improves both the mutual information and the canonical correlation analysis synchrony estimation. Audio–visual synchrony measures aim at highlighting the correlation between speech acoustics and facial movements due to speech production. However head rotation unrelated to speech production may result in spurious correlations. This effect can be mitigated considering vertical motion components only. The best audio–visual synchrony measures combined with MFCCs attained a 20% relative improvement compared to the baseline MFCC only diarisation system. Moreover these novel audio–visual synchrony features provided the best diarisation performances (24.7% DER), outperforming both head pose likelihoods and motion intensity features [6].

Acknowledgements This work was supported by the NCCR Interactive Multimodal Information Management project (IM2). We thank Dr. Gerald Friedland for the Speech/Non-Speech detection output, Dr. Sileye Ba and Dr. Jean-Marc Odobez for the automatic head pose estimation.

9. References

[1] S. Tranter and D. Reynolds, "An Overview of Automatic Speaker Diarization Systems," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, 2006.

[2] C. Wooters and M. Huijbregts, "The ICSI RT07s Speaker Diarization System," *Proc. Rich Transcription Spring Meeting Recognition Evaluation*, 2007.

[3] A. K. Noulas, G. Englebienne, and B. J. A. Krose, "Multimodal speaker diarisation," `www.noulo.net/docs/pubs/noulas09.pdf`, 2009.

[4] K. Otsuka et al., "A Realtime Multimodal System for Analysing Group Meetings by Combining Face Pose Tracking and Speaker Diarisation," in *Proc. ICMI*, 2008.

[5] G. Friedland, H. Hung, and C. Yeo, "Multi-Modal Speaker Diarization of Real-World Meetings using Compressed Domain Video Features," in *Proc. ICASSP*, 2009.

[6] G. Garau and H. Bourlard, "Using audio and visual cues for speaker diarisation initialisation," in *Proc. ICASSP*, 2010.

[7] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal tract and facial behaviour," *Speech Communication*, vol. 26, pp. 23–43, 1998.

[8] J. Hershey and J. Movellan, "Audio–Vision: Using Audio–Visual Synchrony to Locate Sound," in *Proc. NIPS*, 1999, pp. 813–819.

[9] H.J. Nock, G. Iyengar, and C. Neti, "Speaker localisation using audio-visual synchrony: An empirical study," in *Proc. of ACM CIVR*, 2003.

[10] M. Slaney and M. Covell, "FaceSync: a linear operator for measuring synchronization of visual facial images and audio tracks," in *Proc. NIPS*, 2000.

[11] J. Shi and C. Tomasi, "Good Features to Track," in *Proc. IEEE Computer Vision and Pattern Recognition*, 1994.

[12] E. Padilha and J. Carletta, "Nonverbal Behaviours Improving a Simulation of Small Group Discussion," in *Proc. of the 1st Nordic Symposium on Multimodal Communications*, 2003.

[13] J. Carletta et al., "The AMI Meeting Corpus: A Pre-Announcement," *Proc. MLMI*, 2005.

[14] X. Anguera, C. Wooters, and J. Hernando, "Speaker diarization for multi-party meetings using acoustic fusion," in *Proc. ASRU*, 2005.

[15] S.O. Ba, H. Hung, and J.-M. Odobez, "Visual Activity Context for Focus of Attention Estimation in Dynamic Meetings," in *Proc. of ICME*, 2009.

[16] H. Hotelling, "Relations between Two Sets of Variates," *Proc. Biometrika*, vol. 28, pp. 312–377, 1936.

[17] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Transactions on Multimedia*, vol. 9, no. 7, 2007.