Unsupervised Detection of Video Sub-scenes

George Kamberov[†], Matt Burlick[*], Lazaros Karydas[*], Olga Koteogou[*],
[†]Department of Mathematics
University of Alaska, Anchorage, AK 99508
Email: gkamberov@uaa.alaska.edu
[*]Department of Computer Science
Stevens Institute of Technology, Hoboken, NJ 07030
Email: mburlick, lkarydas, okoteoglou @stevens.edu

***Abstract*—The analysis of videos taken by active operators recording human interactions and activities in the field presents a new set of challenges. For brevity in this paper we will call such subject centric field grade videos ad hoc videos of events. Human test subjects readily segment ad hoc videos of events into scene-like segments. These segmentations can not be replicated by the state of the art automatic video segmentation algorithms. We propose and evaluate a method to segment ad hoc videos of events into atomic semantics units. Motivated by [Bel74] we call these units sub-scenes. Our experiments show that the segments detected by human subjects are sequences of sub-scenes. Thus the sub-scenes appear to be a semantic version of the video shots that are used to piece together scenes by state of the art video segmentation algorithms.**

I. Introduction

The goal of this work is to develop methods for analysis and segmentation of ad-hoc videos of events, i.e., videos taken by human operators recording human interactions and activities in the field. The analysis of ad-hoc videos of events is an important growing area of interest, e.g., [DAR11]. These videos are taken by active operators covering scenes involving persons, activities, and interactions the operators deem important. Humans readily segment ad hoc videos of events into scene-like segments. We will call these scene-like segments events. The state of the art video segmentation algorithms [SC00], [CLG09], [VF07], [HXL+11], [RS05], [YYL98], [WCW07], [TCH11], [ZS05], [WTY+08], [BK10], [SMKK12] fail to detect these scene-like segments because they can only detect segments which are collections of consequent video shots while the events detected by the humans rarely are unions of video shots. Table I and Figure 1 illustrate this point. Figure 1 shows the relationship between the shot boundaries detected by a traditional state of the art shot detector and the event boundaries detected by human subjects. In this particular example, which corresponds to video V1 in Table I: **(i)** the shot detector detected two shot boundaries (a hard cut at frame 797 and another transition at frame 798, the later is a false positive shot boundary detection due to an exaggerated object/camera motion); while **(ii)** the crowd workers reported at least five event-like boundaries inside the video (marked by blue vertical lines in the "consensus boundaries" portion of Sub-figure 1a.) which segment the video in at least six pieces (HDSS1 through HDSS6) characterized by the different groups of people that appear to be the focus of operator's attention.

Cutting et al. , [CBC12] show that humans rely extensively on the entrances and exits of characters to mark event boundaries. The importance of the units punctuated by such character exits and entrances, called sub-scenes in [CBC12] is well established in film theory [Bel74]. More generally humans appear to use the change of perceived relative importance of characters to mark event boundaries.

In this paper we present an approach to define and detect sub-scenes boundaries in ad hoc videos of events marked by the change of the perceived group of relative importance. Our goal is to develop an automatic method to detect such sub-scenes. The solution of this problem requires a simultaneous decision who are important enough people and the time periods during which they are important. Our method centers on exploiting the importance cues encoded in the frame compositions by the person taking the ad-hoc video. In Section II we present automatic methods to measure the importance of individuals and define an approach to exploit such measures to detect sub-scenes using the cues provided by the operator. In Section III we describe experiments aimed at explaining the relationship between video events (scenes) and our notion of a sub-scene. The experimental data indicates that similarly to video shots the sub-scenes boundaries are aligned with event boundaries.

The proposed approach can be useful only if we have robust methods to detect and track the participants in ad hoc videos of events. Without excessive manual tunning, the of state of the art human trackers have low accuracy on ad hoc videos of events. One possible approach is to use an appearance frequency-based participation measure as the one defined in (2) and a high accuracy person detection and "tagging" tools like SCAR [KBKK12] (without any tunning the approach achieves 76% accuracy at 83% precision on the tested videos) but SCAR can only tell us whether a person is present in a frame or not. A more informative approach would be to exploit the importance cues encoded in the frame compositions. We define such participation measure in (5) – it requires accurate face position and size information.

II. Participation and sub-scenes

The key to our approach is a method to segment the participants in a video into foreground individuals layer and a background individuals layer. The changes of these foreground groups punctuate the video and segment it into a sequence of sub-scenes.

Throughout this paper we use the notation $o \in f$ whenever an individual o is present in a given video frame f. In particular, for any collection of video frames $\mathcal{V}'$ we will denote

	V1	V2	V3	V4	V5	V6
# frames	1801	1194	1925	3927	294	261
# shot boundaries	2	72	0	1	1	0
# event boundaries	5	12	19	18	5	3

TABLE I. ANALYSIS OF SIX SHORT AD-HOC VIDEOS: CROWD-SOURCED LABELING SHOWS THAT HUMANS READILY SEGMENT VIDEOS INVOLVING HUMAN INTERACTIONS IN EVENTS WITH BOUNDARIES WHICH CAN NOT BE EXPLAINED WITH THE SHOT BOUNDARIES DETECTED BY AN OFF-THE-SHELF SHOT DETECTOR [MAT]

by $\phi_0(\mathcal{V}')$ the set of all actors initially determined to be present in any of the frames of $\mathcal{V}'$:

$$\phi_0(\mathcal{V}') = \{o : \exists \text{ a frame } f \in \mathcal{V}' \text{ s.t. } o \in f\}. \quad (1)$$

A. Defining Participation Probabilities

To analyze the significance of individuals we will need to measure each individual's level of participation. We will represent individual levels of participation as participation probabilities. For example, the participation probability $\mathrm{P}(o; \phi_0(\mathcal{V}'))$ models how important is the individual o relative to all other individuals observed in the collection of frames $\mathcal{V}'$. More generally, if the individual o is a member of a subgroup $\phi(\mathcal{V}') \subset \phi_0(\mathcal{V}')$ then $\mathrm{P}(o; \phi(\mathcal{V}'))$ will denote her participation probability but computed relative to the group $\phi(\mathcal{V}')$.

The simplest method to measure the participation probability of an actor o in a sequence of video frames $\mathcal{V}'$ is to use her appearance frequency. Let $\mathcal{V}'$ be a collection of frames from a video $\mathcal{V}$ and $\phi(\mathcal{V}') \subset \phi_0(\mathcal{V}')$. For every $o \in \phi(\mathcal{V}')$, the frequency-based participation probability is defined as

$$\mathrm{P}(o; \phi(\mathcal{V}')) = \frac{\#\{f \in \mathcal{V}' | o \in f\}}{\sum_{o' \in \phi(\mathcal{V}')} \#\{f \in \mathcal{V}' | o' \in f\}}. \quad (2)$$

This participation measure is very useful for videos on which the existing automatic human trackers do not work. In this case one can work with an automatic high accuracy engine like [KBKK12] since it gives a good method to detect whether a person is present in a frame or not.

We now propose participation probabilities which exploit the importance cues hidden in the frame composition. We call such a participation probability the operator assigned participation measure (OAPM).

The definition of an OAPM is based on the principle that every operator will try (consciously or subconsciously) to compose the frames so that the important individuals are covered in a satisfactory and usually revealing manner.

In a video shot and/or edited by a formally trained operator one can measure the OAPM by exploiting formal frame composition rules as the: the rule of thirds; the type of cropping and shot distances selections; the 180 degree rule; video lighting rules; scene matching of look-position-movement rules, etc. These rules are rarely satisfied in ad-hoc videos taken with a hand held camera in dynamic, cluttered scenarios by operators that may or may not be aware of the formal cinematographic rules. Still, an active operator trying to record a scene is likely to at least keep the persons of interest within many frames, relatively close to the center of the frame (at least to avoid missing important parts of the action), and will strive to have reasonably good views of the subjects. These heuristics lead to a rudimentary and intuitive OAPM. Let $\mathcal{V}'$ be a collection of frames from a video $\mathcal{V}$ and $\phi(\mathcal{V}') \subset \phi_0(\mathcal{V}')$ and for every frame $f \in \mathcal{V}'$ let

$$\phi(\mathcal{V}'; f) = \{o \in \phi(\mathcal{V}') : o \in f\}$$

Let $\mathrm{face}(o, f)$ denote the segment of image pixels occupied by the person o's face in the frame f. For every $o \in \phi(\mathcal{V}'; f)$, the relative participation (within the subgroup $\phi(\mathcal{V}')$) in the frame f can expressed in terms of the of the weights

$$w_d(o; \phi(\mathcal{V}'; f)) = \frac{e^{-\mathrm{dist}(\mathrm{face}(o,f), \mathrm{center}(f))}}{\sum_{o' \in \phi(\mathcal{V}'; f)} e^{-\mathrm{dist}(\mathrm{face}\ (o', f), \mathrm{center}(f))}}$$

and

$$w_a(o; \phi(\mathcal{V}'; f)) = \frac{\mathrm{Area}(\mathrm{face}(o, f))}{\sum_{o' \in \phi(\mathcal{V}'; f)} \mathrm{Area}(\mathrm{face}(o', f))}.$$

The weight $w_d(o; \phi(\mathcal{V}'; f))$ measures the centrality of the face inside the frame f, while the weight $w_a(o; \phi(\mathcal{V}'; f))$ is a cue to the type of camera shot (e.g., wide, or some type of a close up). The area weight is motivated by the heuristics that an operator will try to keep the core characters into something resembling medium or close up shots. (While this is natural expect, it is actually harder to achieve during extended ad hoc video recordings.) The total centrality weight and area weights of a person are defined by

$$w_d(o; \phi(\mathcal{V}') = \sum_{f \in \mathcal{V}'} w_d(o; \phi(\mathcal{V}'; f)). \quad (3)$$

$$w_a(o; \phi(\mathcal{V}') = \sum_{f \in \mathcal{V}'} w_a(o; \phi(\mathcal{V}'; f)). \quad (4)$$

Let us define, the total weight of person $o \in \phi(\mathcal{V}')$ in the frames $\mathcal{V}'$ as

$$w(o; \phi(\mathcal{V}')) = \sum_{f \in \mathcal{V}'} w_d(o; \phi(\mathcal{V}'; f)) w_a(o; \phi(\mathcal{V}'; f))$$

The centrality and shot distance OAPM of o, relative to the group $\phi(\mathcal{V}')$, is defined as

$$\mathrm{P}(o; \phi(\mathcal{V}')) = \frac{w(o; \phi(\mathcal{V}'))}{\sum_{o' \in \phi(\mathcal{V}')} w(o'; \phi(\mathcal{V}'))}. \quad (5)$$

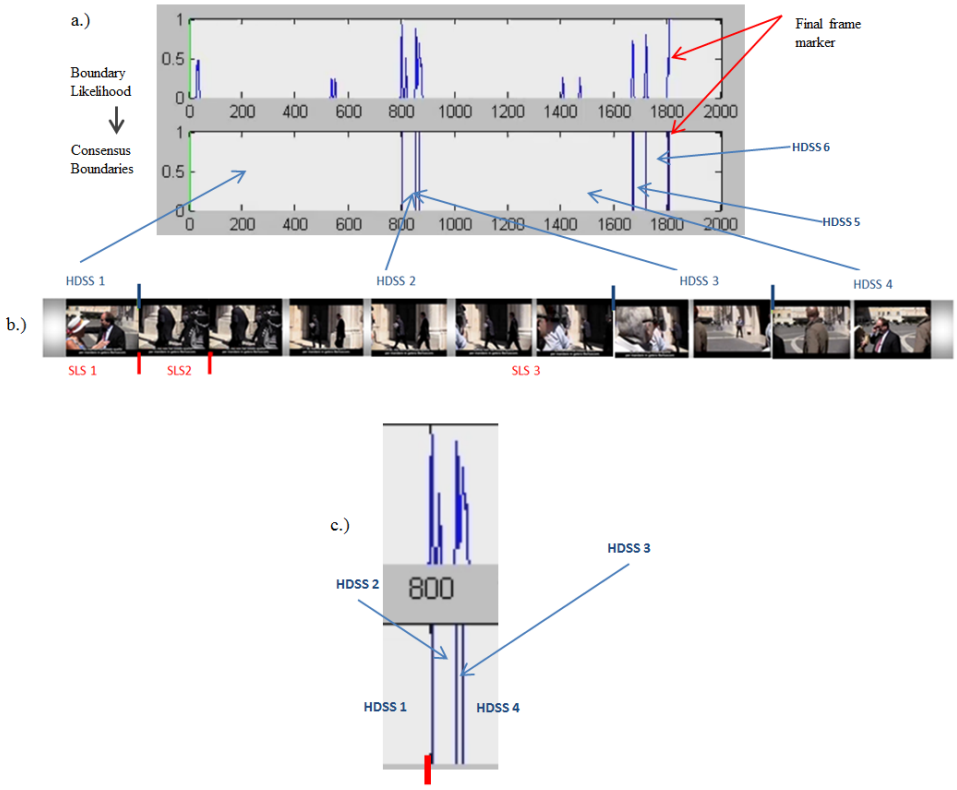

Fig. 1. Sub-figure 1a.) (top) illustrates the likelihood (the ordinate value) that an event-like boundary is detected by the crowd sourced workers at a particular frame (the frame numbers are marked along the abscissa axis) ; Sub-figure 1a.) (bottom) shows the consensus high probability boundaries. In this particular example, which corresponds to video V1 in Table I, the shot detector detected two shot boundaries (a hard cut at frame 797 and another transition at frame 798, the later can be seen as false positive because it is due to an exaggerated object/camera motion). Sub-figure 1b.) shows the frames $(795, 868)$ portion of the time-line of the video. The shot boundaries are marked with red tick-marks at the lower left corner of the frames; the consensus human-detected event-like boundaries extracted from the likelihood function in sub-figure 1a.) are marked with blue tick-marks at the top left corner of the frames. Thus humans perceive four different events/scenes all characterized by the operator's changing focus onto different groups of people and activities/interactions: HDSS1 – an interaction between a lady and a gentleman; HDSS2 – the operator focuses on the a group of people walking into the building; HDSS3 – the operator again changes her focus onto a new group of interest (achieved by panning the camera away from the building entrance); HDSS4 – the focus of interest moves onto the interactions of a group of people including the lady and the gentleman from HDSS1. Depending on training and tuning, the semantic analysis based on the video shots will yield either three distinct scenes (marked by SLS1, SLS2, and SLS3) or only two scenes (SLS1 and the union of SLS2 and SLS3). The top portion of sub-figure 1c.) shows a zoom in on the time-line in sub-figure a.) and the bottom portion shows the consensus crowd-sourced boundaries – the blue verticals lines and the resulting semantic segments HDSS1, HDSS2, HDSS3, and HDSS4 (the red tick-mark shows the detected shot boundaries).

B. *Foreground participation layer*

We are now ready to turn to the definition of a sub-scene, as a continuous sequence of frames describing the interactions of a group of core individuals (a common foreground layer). Following [CBC12] a change in the core group is equivalent to the beginning of a new sub-scene.

The members of the core group in any sequence of frames in an ad hoc video have the top participation probabilities in this sequence. Furthermore, since all of them are equally important to the operator, their participation probabilities should be almost uniform, or equivalently, the roughness of the distribution [HT01] of the participation probabilities of the core group members should be as small as possible. These observations motivate our definitions and approach described below.

The extraction of the core group of individuals which forms the foreground layer covered by a collection of frames is an iterative process during which low participation individuals are peeled off the group until only the core is left.

Given a video $\mathcal{V}$ and an initial collection of persons present $\phi_0(\mathcal{V})$, for any collection of frames $\mathcal{V}' \subset \mathcal{V}$ in the video $\mathcal{V}$ we will extract a hierarchal structure of pre-foreground society layers

$$\phi_0(\mathcal{V}') \supset \phi_1(\mathcal{V}') \supset \ldots \supset \alpha(\mathcal{V}') = \phi_{n(\mathcal{V}')}(\mathcal{V}'),$$

where we will denote by $\phi_{n(\mathcal{V}')}(\mathcal{V}')$ the foreground layer to be uncovered, see (12). For every $j = 0, 1, \ldots, n(\mathcal{V}') - 1$, the intermediate layer $\phi_{j+1}(\mathcal{V}') \subset \phi_j(\mathcal{V}')$ is built by selecting the individuals in $\phi_j(\mathcal{V}')$ whose level of participation is sufficiently high. Instead of trying to learn a cut off threshold we adopt ideas from quantum mechanics and the theory of

random graphs [FDBV01], [HT01], [Zyc03], [ABG+05]. We model actor's participation levels as the events admitted by a stochastic variable and following [Zyc03] we can now model with

$$\pi(\phi_j(\mathcal{V}')) = 1/\sum_{o\in\phi_j(\mathcal{V}')} \mathrm{P}(o;\phi_j(\mathcal{V}'))^2 \quad (6)$$

the effective number of different states. Therefore we postulate that the j-th foreground layer contains no more than $\mathrm{int}(\pi(\phi_j(\mathcal{V}')))$ actors, namely the actors with the top $\mathrm{int}(\pi(\phi_j(\mathcal{V}')))$ participation probabilities. Therefore,

$$|\phi_{j+1}(\mathcal{V}')| \leq \mathrm{int}(\pi(\phi_j(\mathcal{V}'))) \quad (7)$$

and the members of $\phi_{j+1}(\mathcal{V}')$ are chosen so that

$$\min_{o\in\phi_{j+1}(\mathcal{V}')} \mathrm{P}(o;\phi_j(\mathcal{V}')) > \max_{o\in\phi_j(\mathcal{V}')\setminus\phi_{j+1}(\mathcal{V}')} \mathrm{P}(o;\phi_j(\mathcal{V}')). \quad (8)$$

By construction, $\phi_{j+1}(\mathcal{V}') \subset \phi_j(\mathcal{V}')$,

$$\mathrm{P}(o;\phi_{j+1}(\mathcal{V}')) \geq \mathrm{P}(o;\phi_j(\mathcal{V}')),\ \forall o \in \phi_{j+1}(\mathcal{V}'). \quad (9)$$

The hierarchy is constructed so that low participation individuals are weeded out in an unsupervised manner and thus:

$$\pi(\phi_0(\mathcal{V}')) < \pi(\phi_1(\mathcal{V}')) < \ldots < \pi(\phi_{n(\mathcal{V})}(\mathcal{V}')). \quad (10)$$

We call $\mathrm{IP}(\phi_j(\mathcal{V}')) = \pi(\phi_j(\mathcal{V}'))$ the index of overall participation by the members of the layer $\phi_j(\mathcal{V}')$.

To identify the foreground layer for the collection of frames $\mathcal{V}'$ let us consider the behavior of the roughness of each distribution $\mathrm{P}(o;\phi_j(\mathcal{V}')), j = 0, 1, \ldots, n(\mathcal{V}')$:

$$MR(\phi_j(\mathcal{V}')) = \sum_{o\in\phi_j(\mathcal{V}')} \left(\mathrm{P}(o;\phi_j(\mathcal{V}')) - \frac{1}{|\phi_j(\mathcal{V}')|}\right)^2. \quad (11)$$

We now propose a **criterion** to identify the foreground (core)

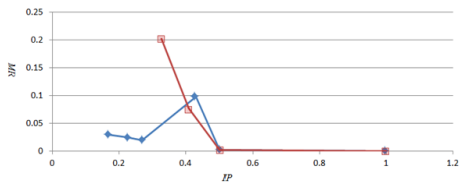

Fig. 2. Two IP/MR graphs obtained by peeling off low participation individuals in two different videos. The Blue video involves 7 individuals, while the Red video involves a group of 8 individuals. The local minimum in the blue curve happens precisely when the core group is split from the rest of the video population. Manual inspection and the operators notes indicate that the Blue video is about these two people passing through several short scenes, each of which contains a few other low participating (transient) individuals. The lack of local extrema in the Red curve and its steady and initially rapid descend are evidence that the video operator is attempting to cover several individuals in a small number of settings (two settings were identified by independent manual inspection) connected by the thread of a single high persistence individual.

people in a collection of frames based on the properties of the IP/MR graph (Figure 2, shows the IP/MR graphs for two different videos.):

Consider the participation vs. roughness curve

$$[(\pi(\phi_j(\mathcal{V}')), MR(\phi_j(\mathcal{V}'))),\quad j = 1, \ldots, n(\mathcal{V}')].$$

Let

$$j_* = \min_j\{j|\, MR(\phi_{j+1}(\mathcal{V}')) > MR(\phi_j(\mathcal{V}'))\}.$$

Then by definition the foreground layer for the collection of frames $\mathcal{V}'$ is

$$\alpha(\mathcal{V}') \stackrel{\mathrm{def}}{=} \phi_{j_*}(\mathcal{V}'). \quad (12)$$

Note that the first local minimum value of the roughness in the IP/MR graph is $MR\,(\phi_{j_*}(\mathcal{V}'))$.

The definition of a foreground layer is based on the following heuristics:

[h1] At any given time interval the operator has identified a core group of interest. Thus the members of the core group are more important (have higher participation probabilities) than the rest of the characters present in this video segment and all core members are almost equally important locally in time.

[h2] In order to keep track of the unfolding events the video operator will try to react as rapidly as possible to keep covering the people that he/she considers to be currently the core individuals.

C. Sub-scenes

We are now ready to turn to the definition of a sub-scene, as a continuous sequence of frames describing the interactions of a group of core individuals (a common foreground layer). Following [CBC12] a change in the core group is equivalent to the beginning of a new sub-scene.

Definition 5: A sub-scene in a video $\mathcal{V}$ is a triple

$$(\mathcal{I}, \alpha(\mathcal{I}), \beta(\mathcal{I}) = \phi_0(\mathcal{I}) \setminus \alpha(\mathcal{I})),$$

where

a.) $\mathcal{I} \subset \mathcal{V}$ is an uninterrupted sequence of video frames;
b.) $\alpha(\mathcal{I}) \subset \phi_0(\mathcal{I})$ is the foreground layer for $\mathcal{I}$;
c.) $o \in f, \forall o \in \alpha(\mathcal{I}), \forall f \in \mathcal{I}$;
d.) The scene is of maximal time length with respect to inclusion for the fixed foreground layer.

The last condition in the definition implies that $\mathcal{I} = \cup\mathcal{I}'$ over all triples $(\mathcal{I}', \alpha(\mathcal{I}'), \beta(\mathcal{I}'))$ such that $\alpha(\mathcal{I}) = \alpha(\mathcal{I}')$ and $\mathcal{I}' \cap \mathcal{I} \neq \emptyset$.

The **subScenes** algorithm, see Algorithm 1, identifies a group of foreground characters in a video (the overall foreground layer) and then segments the video into maximal length sub-scenes.

III. Sub-scenes and Event Boundaries

To test how well the sub-scenes we defined here match event boundaries we analyzed six videos denoted in this paper by V1 (1801 frames, involving 8 participants), V2 (1194 frames, involving 19 participants), V3 (1925 frames, involving 4 participants), V4 (3927 frames, involving 10 participants), V5 (294 frames, involving 14 participants), and V6 (261 frames, involving 9 participants).

We used crowd sourcing to segment the videos into (ground truth) events. To measure the inherent fuzziness of the event

Algorithm 1 subScenes Computes the list of sub-scenes captured in a sequence of video frames $\mathcal{I}$

Input: 1. A sequence of video frames $\mathcal{I}$
Input: 2. A list of characters $\phi_0(\mathcal{I})$ in $\mathcal{I}$
Output: The list of sub-scenes subScenes
  subScenes $\leftarrow \emptyset$
  $\alpha(\mathcal{I}) \leftarrow$ Foreground Characters in $\mathcal{I}$ among $\phi_0(\mathcal{I})$
  if $(\mathcal{I}, \alpha(\mathcal{I}), \phi_0(\mathcal{I}) \setminus \alpha(\mathcal{I}))) ==$ sub-scene **then**
    return append(subScenes, $(\mathcal{I}, \alpha(\mathcal{I}), \phi_0(\mathcal{I}) \setminus \alpha(\mathcal{I})))$
  else
    $\mathcal{I}' \leftarrow \{f | f$is a frame in $\mathcal{I}$ s.t. $\exists o \in \alpha(\mathcal{I})$ present in $f\}$
    $\mathcal{I}'' \leftarrow \mathcal{I} \setminus \mathcal{I}'$
    $\mathrm{C}(\mathcal{I}') \leftarrow$ minimal covering of $\mathcal{I}'$ by sequences of consecutive frames s.t., $\forall [s, e] \in \mathrm{C}(\mathcal{I}')$, $\forall o \in \phi_0(\mathcal{V})$, if $\exists f' \in [s, e]$, s.t. $o \in f'$ then $o \in f, \forall f \in [s, e]$
    $\mathrm{C}(\mathcal{I}'') \leftarrow$ minimal covering of $\mathcal{I}''$ by sequences of consecutive frames
    while $\mathcal{J} \in \mathrm{C}(\mathcal{I}') \cup \mathrm{C}(\mathcal{I}")$ **do**
      $\phi_0(\mathcal{J}) = \{o \in \phi_0(\mathcal{I}) | \exists f \in \mathcal{J}$s.t.$o \in f\}$
      return append(subScenes, subScenes$(\mathcal{J}, \phi_0(\mathcal{J})))$
    end while
  end if

boundaries we asked the labelers to mark both the most likely locations of scene boundaries and the uncertainty intervals for the location for each event transition boundary they detected. This data was used to build an event boundary likelihood for each video and to extract ground truth video segmentation for each video. The videos V1 and V3 are included in the supplement, together with the respective: ground truth event likelihoods, consensus event boundaries, sub-scenes boundaries extracted using the OAMP participation measure. For each video we computed the median uncertainty for the event boundary locations over all boundaries (in this video) and subjects, see the bottom row in Table III. This median uncertainty gives us a measure of the inherent error in the ground truth event boundary locations.

We build a high accuracy perfect precision semi-automatic labeling tool using SCAR and Google's Picasa. We used this labeling tool and the ground truth data to test:

First, how well our method for extraction of sub-scenes would work if we had an almost perfect tracker? We computed the OAPM, and then extracted the sub-scenes. We then compared how well these sub-scenes boundaries align with the ground truth segmentations of the video. The results in Table III row 3 show that the automatically detected sub-scenes boundaries align very well with ground truth scene boundaries given the fuzziness of the ground truth.

Second, how much do we loose if instead of almost perfect tracking we have an automatic high accuracy and high precision detection method that can be used to establish the presence or absence of a person in each video frame? We computed the appearance frequency-based sub-scene boundaries. The results in Table III row 2 show that the frequency-based boundaries also align well with the ground truth. See also Figure 3 for a comparison of alignment of the sub-scenes boundaries obtained using the frequency-based measure and the OAPM for video V1. Thus given a good presence detector we can use it to detect sub-scenes.

Figure 4 shows the performance of the sub-scene detector using a frequency-based participation measure on relatively noisy but automatically generated SCAR data.

The results in Table II show that: (i) The state of the art video segmentation methods can not be used to replicate the human segmentations - because they can only detect events that are sequences of video shots and so they will fail when the number of video shots is smaller than the number of events. (ii) Our approach detects a sufficient number of sub-scenes to replicate the human segmentations. The results in Table III and Figure 3 show that the sub-scenes can be grouped to match the events detected by the human labelers.

IV. Conclusions

A new set of methods is needed perform event segmentation of ad-hoc videos taken by human operators recording human interactions and activities in the field. In this paper we presented a method to detect sub-scenes exploiting the importance cues provided by the scene composition. The method is completely automatic. The evaluation shows that the sub-scene boundaries align well with manually extracted event boundaries. Thus the sub-scenes appear to be a semantic version of traditional video shots and can be used to piece together scenes by state of the art video segmentation algorithms when video shots extraction is uninformative or unreliable.

Acknowledgments

This work is partially supported by NSF Award Number 0916610.

References

[ABG+05] C Aubin, C Bernard, Steven Gottlieb, EB Gregory, Urs M Heller, JE Hetrick, J Osborn, R Sugar, Ph de Forcrand, and O Jahn. The scaling dimension of low lying dirac eigenmodes and of the topological charge density. *Nuclear Physics B-Proceedings Supplements*, 140:626–628, 2005. 4

[Bel74] R. Bellour. The obvious and the code. *Screen*, 15:7–17, Winter 1974. 1

[BK10] N. Brindha and C. Kalaiarasan. Certain investigations on video scene segementaion techniques. In *Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference on*, pages 1–4, 2010. 1

[CBC12] J. E. Cutting, K. L. Brunick, and A. Candan. Perceiving event dynamics and parsing hollywood films. *Journal of Experimental Psychology: Human Perception and Performance*, 2012. Online First Publication, March 26. 1, 3, 4

[CLG09] V.T. Chasanis, A.C. Likas, and N.P. Galatsanos. Scene detection in videos using shot clustering and sequence alignment. *Multimedia, IEEE Transactions on*, 11(1):89 –100, jan. 2009. 1

[DAR11] I2O DARPA. Visual media reasoning (vmr). DARPA Broad Agency Announcement, 2011. 1

[FDBV01] Illes J Farkas, Imre Derenyi, Albert-Laszlo Barabasi, and Tamas Vicsek. Spectra of real-world graphs: Beyond the semicircle law. *Physical Review E*, 64(2):026704, 2001. 4

[HT01] Peter Harremoës and F Topsoe. Inequalities between entropy and index of coincidence derived from information diagrams. *Information Theory, IEEE Transactions on*, 47(7):2944–2960, 2001. 3, 4

[HXL+11] Weiming Hu, Nianhua Xie, Li Li, Xianglin Zeng, and S. Maybank. A survey on visual content-based video indexing and retrieval. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 41(6):797 –819, 2011. 1

	V1	V2	V3	V4	V5	V6
# frames	1801	1194	1925	3927	294	261
# events (ground truth)	6	13	20	19	6	4
# video shots	3	73	1	2	2	1
# sub-scenes using the participation measure defined in (2)	19	48	42	32	6	7
# sub-scenes using the participation measure defined in (5)	12	53	93	33	6	11

TABLE II. ANALYSIS OF SIX SHORT AD-HOC VIDEOS: ROW 3 SHOWS THE NUMBER OF EVENTS IN EACH VIDEO OBTAINED BY CROWD SOURCING; ROW 4 SHOWS THE NUMBER OF VIDEO SHOTS DETECTED IN EACH VIDEO BY AN OFF-THE-SHELF SHOT DETECTOR [MAT]; ROWS 5 AND 6, SHOW THE NUMBER OF SUB-SCENES DEFINED IN SECTION II-C AND DETECTED BY THE **subScene** ALGORITHM, (SEE **Algorithm 1**).

	V1	V2	V3	V4	V5	V6
subScenes using the participation measure defined in (2)	0	0	7	19	2	1
subScenes using the participation measure defined in (5)	0	0	6	6	2	1
Median subjective boundary uncertainty	2	2	6	6	2	2

TABLE III. THE MEDIAN MISS-ALIGNMENT (MEASURED IN NUMBER OF FRAMES) BETWEEN GROUND TRUTH EVENT BOUNDARY LOCATIONS AND THE LOCATIONS OF SUB-SCENE BOUNDARIES OBTAINED BY THE **subSbcenes** ALGORITHM .

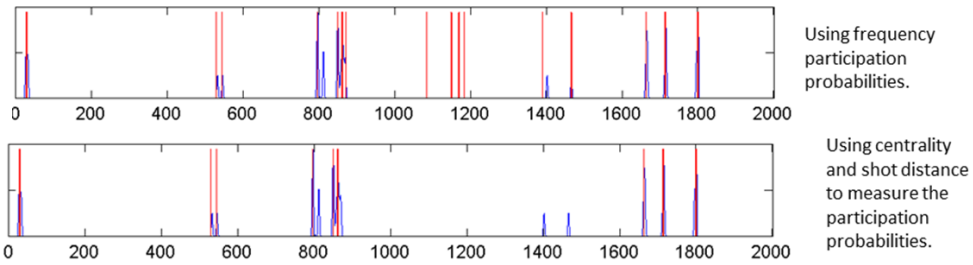

Fig. 3. The detected sub-scene boundaries in video V1 (marked by the red vertical lines) using the frequency -based participation measure and the centrality and shot distance OAPM. The ground truth event boundary likelihood is the blue curve.

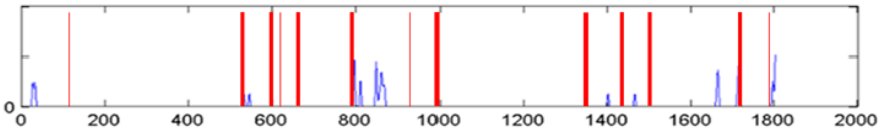

Fig. 4. The detected sub-scene boundaries in video V1 (marked by the red vertical lines) using the frequency -based participation on on SCAR data [KBKK12]. The ground truth event boundary likelihood is the blue curve.

[KBKK12] G. Kamberov, . Burlick, M, L. Karydas, and O. Koteogou. Scar: Dynamic adaptation for person detection and persistence analysis in unconstrained videos. In *Proceedings of the 8th International Symposium Visual Computing, ISVC2012*, volume 7432. Springer Lecture Notes in Computer Science, 2012. 1, 2, 6

[Mat] Johan Mathe. Shotdetect. http://johmathe.name/shotdetect.html. 2, 6

[RS05] Z. Rasheed and M. Shah. Detection and representation of scenes in videos. *Multimedia, IEEE Transactions on*, 7(6):1097 – 1105, dec. 2005. 1

[SC00] H. Sundaram and Shih-Fu Chang. Video scene segmentation using video and audio features. In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, volume 2, pages 1145–1148 vol.2, 2000. 1

[SMKK12] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, and J. Kittler. Differential edit distance: A metric for scene segmentation evaluation. *Circuits and Systems for Video Technology, IEEE Transactions on*, 22(6):904–914, 2012. 1

[TCH11] Tsung-Hung Tsai, Wen-Huang Cheng, and Yung-Huan Hsieh. Dynamic social network for narrative video analysis. In *Proceedings of the 19th ACM international conference on Multimedia*, MM '11, pages 663–666, New York, NY, USA, 2011. ACM. 1

[VF07] Alessandro Vinciarelli and Sarah Favre. Broadcast news story segmentation using social network analysis and hidden markov models. In *Proceedings of the 15th international conference on Multimedia*, pages 261–264. ACM, 2007. 1

[WCW07] Chung-Yi Weng, Wei-Ta Chu, and Ja-Ling Wu. RoleNet: treat a movie as a small society. In *MIR '07: Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 51–60. ACM, 2007. 1

[WTY+08] Jingdong Wang, Xinmei Tian, Yichen Yang, Zheng-Jun Zha, and Xian-Sheng Hua. Optimized video scene segmentation. In *Multimedia and Expo, 2008 IEEE International Conference on*, pages 301–304, 2008. 1

[YYL98] M. Yeung, B-L. Yeo, and B. Liu. Segmentation of video by clustering and graph analysis. *Computer Vision and Image Understanding*, 71(1):94 – 109, 1998. 1

[ZS05] Y. Zhai and M. Shah. Automatic segmentation of home videos. In *Proc. IEEE ICME 05*, pages 9–12, 2005. 1

[Zyc03] K. Zyczkowski. Rényi extrapolation of Shannon entropy. *Open Systems and Information Dynamics*, 10:297–310, 2003. 4