Concept Based Hybrid Fusion of Multimodal Event Signals

Yuhui Wang*, Christian von der Weth†, Yehong Zhang‡, Kian Hsiang Low‡, Vivek K. Singh§, and Mohan Kankanhalli‡
*NUS Graduate School for Integrative Sciences and Engineering
†SeSaMe Centre, Interactive & Digital Media Institute ‡Department of Computer Science,
National University of Singapore, Singapore
§School of Communication and Information, Rutgers University, New Brunswick, U.S.A.
* wangyuhui@u.nus.edu †vonderweth@nus.edu.sg ‡{yehong, lowkh, mohan}@comp.nus.edu.sg §vivek.k.singh@rutgers.edu

***Abstract*— Recent years have seen a significant increase in the number of sensors and resulting event related sensor data, allowing for a better monitoring and understanding of real-world events and situations. Event-related data come from not only physical sensors (e.g., CCTV cameras, webcams) but also social or microblogging platforms (e.g., Twitter). Given the wide-spread availability of sensors, we observe that sensors of different modalities often independently observe the same events. We argue that fusing multimodal data about an event can be helpful for more accurate detection, localization and detailed description of events of interest. However, multimodal data often include noisy observations, varying information densities and heterogeneous representations, which makes the fusion a challenging task. In this paper, we propose a hybrid fusion approach that takes the spatial and semantic characteristics of sensor signals about events into account. For this, we first adopt the concept of an image-based representation that expresses the situation of particular visual concepts (e.g. "crowdedness", "people marching") called Cmage for both physical and social sensor data. Based on this Cmage representation, we model sparse sensor information using a Gaussian process, fuses multimodal event signals with a Bayesian approach, and incorporates spatial relations between the sensor and social observations. We demonstrate the effectiveness of our approach in as a proof-of-concept over real-world data. Our early results show that the proposed approach can reliably reduce the sensor-related noise, localize event place, improve event detection reliability, and add semantic context so that the fuses data provide a better picture of the observed events or situations.**

***Keywords*-multimodal fusion; situation understanding; multisensor data analysis**

I. Introduction

Detecting real-world events using distributed multimodal sensors is an important research problem with applications in defense, civic processes, sports, entertainment, and transportation. The exponential growth in social media (e.g. Twitter, Weibo, Flickr) and physical sensors (e.g. CCTVs, satellite imagery, traffic sensors) over the last few years have created an unprecedented opportunity to leverage such multimodal data for distributed event detection. These multimodal *event signals* describe different aspects of emerging situations and reflect various spatio-temporal patterns [10]. Fusing and exploring sensor streams that capture different perspectives of an event, could be useful for better accuracy, localization, semantic interpretation of various events [13].

However, the fusion of such multimodal data remains a challenge due to the heterogeneous data representations, different information densities, and the inherent noise in each modality. Thus there is a need for not only a unified data representation format, but also a sophisticated framework that can combine and analyze such multimodal data for better event detection. Building upon a recent effort in multimedia research that has defined E-mages [12] as a unified representation for spatio-temporal data, in this work, we adopt evolving (image-like) spatial grid representation to capture heterogeneous event data. Such representation provides a generic way to model heterogeneous spatio-temporal data and also allows for the use of a rich repository of image processing algorithms (e.g. convolution) to easily derive semantically useful event information from such data. From a human user perspective, image, as an artifact that depicts or records visual perception, is also an intuitive way to visualize and understand different phenomena. Building further along this line of work, we propose to use the E-mage representation for designing a hybrid fusion framework for heterogeneous event signals. While our approach is generic, we focus on two important modalities - physical sensors (CCTV cameras) and social (Twitter tweets), which are utilized to generate *sensor Cmage* and *social Cmage* respectively - to ground the discussion.

A sensor Cmage is generated by aggregating *sensor decisions* from multiple physical sensors based on their spatial distribution. A sensor decision is considered the confidence of specific visual concept extracted from the sensor. For example, a crowded being detected from an image captured by a camera with some confidence x (valued between 0 1); or a semantic word "protest" detected from the social stream as an occurring event with high frequency posted in an area. They represent a sensor confirming some event's occurring. Higher sensor decision values mean higher confidence of detecting such event. The geo-locations of the sensors define the corresponding pixel's positions in the image and the

intensity of each pixel is computed by extracting semantic concepts [6] from camera snapshots. For example, Figure 1 shows how a "crowd" image is generated from Manhattan CCTV camera readings.

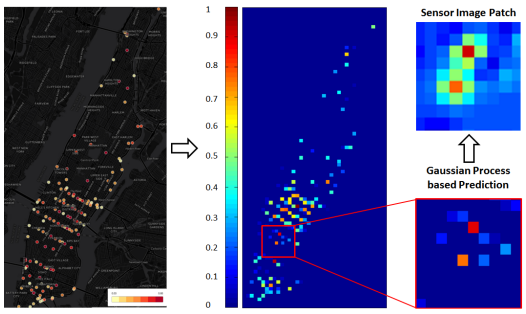


Figure 1: Generating Sensor Cmage from Sensors Map. Left: physical sensors (CCTV Cameras) detecting particular concept ("Crowd") with different probabilities. Middle: corresponding conversion into *sensor Cmage* with sparse pixels. Right: simulating dense sensor readings by applying Gaussian process model to predict missing pixels in a given region.

Due to the intrinsic properties of the sensor and social information, event images usually contain noise to different extents. For example, the location in a social Cmage may be incorrect if people discuss an event at locations other than the event's origin. In addition, the distributed physical sensors have different sparsity compared with the social feeds, which makes event information unavailable at some locations (pixels). Such noise and sparsity properties make event localization and situation understanding a hard problem. To tackle this problem, we design a hybrid fusion framework including Gaussian process model based prediction, Bayesian method based decision fusion as well as a spatial fusion to fuse sensor Cmage and social Cmage, in order to eliminate noise, localize event and uncover its semantic details. The fusion method provides insights for several groups of stakeholders including policy makers, tourists, citizens or city planners.

Thus, our **contributions** in this work include: 1) leveraging multimodal information for better situation understanding; 2) proposing an image-based hybrid fusion method featuring sensor decision and spatial information; 3) reducing noise in sensor data for better event detection, and 4) uncovering event details from the fused images.

II. Related Work

Multimodal sensory systems are widely used to characterize important patterns associated with regional situations. Given a variety of heterogeneous features from data, issues such as how to fuse and which to fuse [18] has drawn much attention. Comprehensive surveys of the existing multimodal fusion techniques for target tracking or object detection in surveillance application can be found in [1]. Recently, large-scale situation awareness and understanding using information from different sources, especially incorporating social media data, has drawn increasing attention. Leveraging data of more diverse social media (i.e., Twitter, Flickr) as well as open data, Kuo et al. [8] mines urban activities in New York City across social media in both visual and semantic perceptions and demonstrate a number of interesting applications revealing patterns related to urban dynamics (e.g., traffic pattern, sentiment, human activities and fashion styles) of NYC. Jou et al. [7] designed a system that extracts "who","what","when" and "where" containing a multimodal perspective from heterogeneous multimedia news sources. Integrating social and sensor data, Pan et al. [10] detects and describes traffic anomalies using human mobility data and social media data. An event is considered as bank-of-concepts [9], fusing multiple social and sensor Cmages for event detection and situation understanding therefore involves exploring relations among semantic concepts. In contrast to existing works in multimodal fusion, however, we consider challenges related to sensor decision, the spatial distribution of event signals.

Image fusion is well studied to combine information from two or more images of a scene into a single composite image so that the fused image is more informative and suitable for either visual perception or computer processing [4]. The objective in image fusion is to 1) reduce uncertainty and minimize redundancy; 2) maximizing relevant information particular to an application or task. A lot of works have been done in image fusion involving issues of multisource [15], multispectrum, multiresolution and multifocus [16] image fusion, where desired properties of original images are extracted and then combined in final fused images in an application-oriented manner. However, different from traditional image fusion problems, event images generated from social or physical sensor data contain pixels of semantic meaning (e.g., concepts or terms of a particular event) rather than photon energies striking at any particular location through light reflection. Since both social or sensor Cmages contain semantic information, it is worth exploring word semantic relatedness in the fusion process. To the best our knowledge, we are the first to address hybrid fusion of physical and social sensors data with different modalities.

III. Proposed Approach

This section describes the formalization and techniques proposed in our hybrid fusion pipeline, shown in Fig. 2. We first extract event signals from both physical sensors and social media; these signals are then mapped to "pixels" in a Cmage based on their corresponding geo-locations and event decisions (how confidence one signal represents a particular concept). After that, Gaussian Process, Bayesian fusion and spatial fusion are then applied sequentially to generate final event Cmage which represent the a situation.

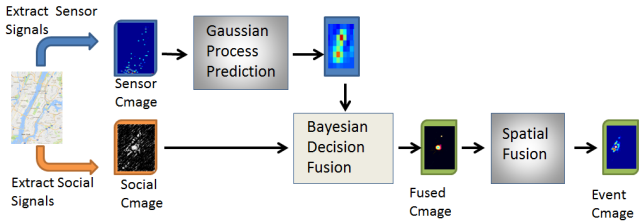


Figure 2: Cmage-based Multimodal Hybrid Fusion Pipeline.

A. From Event Signals to Event Cmages

Sensor Event Cmage: C^{sen} can be generated from a set of M physical sensors $S^{sen} = \{SEN_1, ...SEN_M\}$ in a region bounded by upper-left corner $P_{ul} = (lat_{ul}, lon_{ul})$ and lower-right corner $P_{dr} = (lat_{dr}, lon_{dr})$ in terms of geo-coordinates in the physical world. Each sensor $SEN_m = G_m \times R_m$ is composed of its geo-location $G_m = (lat_m, lon_m)$ and its environment reading R_m (e.g. image captured by a camera, humidity value measured by a weather sensor). A region is then separated into grids, using a user-defined grid size r_{sen}, to form the sensor event Cmage $C^{sen} = [e^{sen}_{ij}]_{H\times W}$, where $H = (lat_{ul} - lat_{dr})/r_{sen}$ and $W = (lon_{ul} - lon_{dr})/r_{sen}$ and sensor pixel $e_{ij} = \mathbb{F}(SEN_m)$; $\mathbb{F}$ is the function that transforms the sensor readings into numeric values, such as a concept detector [6] for an image, a direct copy of air quality index [5], etc., representing the strength of event signal with particular semantic meanings. The mapping from sensor SEN_m location to corresponding Cmage coordinate is defined by $i = \mathbb{LAT}(G_m) = |lat_{ul} - lat_m|/r_{sen}$, $j = \mathbb{LON}(G_m) = |lon_{ul} - lon_m|/r_{sen}$.

Social Event Cmage: C^{soc} is generated from a set of social observations $S^{soc} = \{SOC_1, ...SOC_M\}$ in the same region as physical sensors, where each observation $SOC_m = G_m \times POST_m$ contains its corresponding geo-location and the content $POST_m$ (e.g. the tweet text). We define such posted content $POST = \{term_1, ..., term_c\}$ as a set of terms (or words). Different from pixels of sensor event Cmage where the value of each pixel is derived directly from corresponding one physical sensor, the pixel values in social event Cmage relate to nearby social observations. We propose using two methods to represent the "social pixel" [12]: (1) density based signals; and (2) term frequency based signals. Density based signal method considers the density of nearby posts that contains the particular term. Specifically, given $C^{soc} = [e^{soc}_{ij}]_{H\times W}$ of $term_x$ and a radius r, for social observation SOC_m, $e_{ij} = \mathbb{F}(term_x, SOC_m, r)$; where $\mathbb{F} = \sum_{k=1, dist(G_k, G_m)<r}^{|S^{soc}|} \mathcal{H}(POST_k, term_x)$ is the number of surrounding social observations whose post $POST_k$ contains the term $term_x$, indicated by $\mathcal{H}(POST_k, term_x)$ and $dist(G_k, G_m)$ is the Euclidean distance of two social observation locations. The mapping from social SOC_m location to corresponding pixel coordinate is defined similar to that in sensors event Cmage, but with grid size r_{soc}. For term frequency based method, pixel value e_{ij} is calculated as TF-IDF values defined in [13], which generate each term's weight by considering history posts in the same location. Terms with higher weight mean that they are frequently discussed currently but seldom discussed in the past days.

B. Hybrid Event Image Fusion

1) Sensor Cmage pixel value estimation using noisy and sparse observations: Due to the intrinsic characteristic and sparse spatial distribution of sensors, a sensor Cmage will be generated with many empty and noisy pixels, which causes problem for the later fusion with social Cmages. In particular, fusing social pixel with a false empty pixel (e.g. the one between the two red pixels in bottom-right magnified patch of Fig. 1) will result in an empty pixel reflecting there is no event, which is not the truth. To solve this problem, we assume the sensor readings over an urban area to be realized from a Bayesian non-parametric model, Gaussian process (GP) [14], which incorporates noise model and allows the spatial correlation of sensor readings (sensor pixels) to be formally characterized in terms of their locations in the Cmage. This property enables predicting empty pixels using observed sensor readings. Specifically, assuming a sensor Cmage $C^{sen}_{H\times W}$ is defined as $\{(\boldsymbol{p}_i, e_i)|i = 1, ..., H \times W\}$, where $\boldsymbol{p}_i$ and e_i are the pixel's position ($\boldsymbol{p}_i = [lat_i, lon_i]$) and intensity respectively, we model the joint distribution of Q observed pixel values $\mathbf{e} = [e_1, ..., e_Q]^\top$ and predicted pixel values $\hat{\mathbf{e}} = [e_{Q+1}, ..., e_{W\times H}]^\top$ at the test locations under the prior as:

$$\begin{bmatrix} \mathbf{e} \\ \hat{\mathbf{e}} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(\mathcal{P}, \mathcal{P}) + \sigma_n^2 I & K(\mathcal{P}, \mathcal{P}_\star) \\ K(\mathcal{P}_\star, \mathcal{P}) & K(\mathcal{P}_\star, \mathcal{P}_\star) + \sigma_n^2 I \end{bmatrix}\right) \tag{1}$$

where $\mathcal{P} = [\boldsymbol{p}_1, ..., \boldsymbol{p}_Q]$, and $\mathcal{P}_\star = [\boldsymbol{p}_{Q+1}, ..., \boldsymbol{p}_{H\times W}]$ are the observed and predicted feature matrix respectively; σ_n^2 is the noise variance; and the elements in covariance matrix $K(\cdot, \cdot)$ reflecting correlation between two pixels positions $\boldsymbol{p}_m$ and $\boldsymbol{p}_n$ are defined by the covariance function:

$$k(\boldsymbol{p}_m, \boldsymbol{p}_n) = \sigma_s^2 exp(-\frac{1}{2}\sum_{d=1}^{2}(\frac{\boldsymbol{p}_{m,d} - \boldsymbol{p}_{n,d}}{l_d})^2) \tag{2}$$

where $\boldsymbol{p}_{m,d}(\boldsymbol{p}_{n,d})$ is the d-th component of 2D (lat and lon) vector $\boldsymbol{p}_m$ and $\boldsymbol{p}_n$, and the hyperparameters σ_s^2, l_1, l_2 are signal variance, and length-scales respectively that can be learned using maximum likelihood estimation. Note that term $\boldsymbol{p}_{m,d} - \boldsymbol{p}_{n,d}$ measures the geographic distance of two locations in terms of latitude or longitude.

Having this covariance matrix, values of predictive pixels can be defined by the Gaussian process regression equations:

$$\hat{\mathbf{e}} = k_\star^\top (K + \sigma_n^2 \boldsymbol{I})^{-1} \mathbf{e} \tag{3}$$

where $k_\star$ is the $Q \times (W \times H - Q)$ covariances matrix between predicted pixels and observed pixels, and $K = K(\mathcal{P}, \mathcal{P})$ is

the $Q \times Q$ covariance matrix of observed pixels; $\mathbf{e}$ is $Q \times 1$ observation vector.

2) Event decision fusion: We undertake a pixel-by-pixel fusion between sensor and social Cmages. Specifically, we adopt a Bayesian approach based confidence fusion based on [2], for the sake of simplicity and computationally efficiency. Each pixel e_{ij}^{soc} (e_{ij}^{sen}) represents event decision's confidence from corresponding locations and the fused confidence c_{ij} is computed as:

$$f_{ij} = f(e_{ij}^{soc}, e_{ij}^{sen}) = \frac{e_{ij}^{soc} \cdot e_{ij}^{sen}}{e_{ij}^{soc} \cdot e_{ij}^{sen} + (1 - e_{ij}^{soc})(1 - e_{ij}^{sen})} \tag{4}$$

Using this fusion method, fusing two pixels of high confidence will result in a higher confidence. Fusing high confidence with low confidence pixels (which means conflicting observation) will result in a value close to the lower one and two low confidence pixels will result in a much lower value than either one of them.

3) Spatial fusion: For each social observation or sensor reading, spatial fusion considers its surrounding signals which could contribute to the considered event signal. The range to consider is defined by a fixed reference window $W^{w\_size \times w\_size}$. Windows size is set to be flexible so that users can specify the size of area based on the specific type of the events. Given a reference signal, each signal within the window will be assigned a weight based on their distance to the referenced signal. This is valid since a geographically closer signal has a higher influence. Such a spatial fusion model is then defined by the fusion function F as follows:

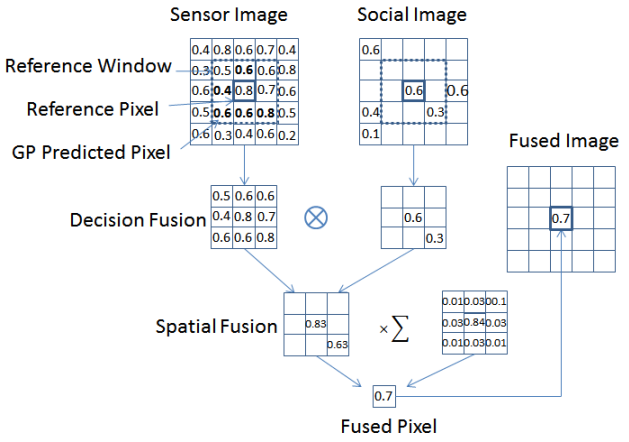


Figure 3: Illustration of Decision and Spatial Fusion

$$F(C^{soc}, C^{sen}, w\_size) = \{e_{ij}^{fus}\} \tag{5}$$

$$e_{ij}^{fus} = \sum_{xy \in W^{ij}} w_{xy} \cdot f_{ij} \tag{6}$$

where $|x - i| \leq \frac{w\_size-1}{2}, |y - j| \leq \frac{w\_size-1}{2}$

$$w_{xy} = \alpha \cdot e^{-\sqrt{(x-i)^2+(y-i)^2}} \tag{7}$$

w_{xy} is the weight given to the neighbouring pixels of referenced pixel e_{ij}^{fus} based on their distances to it. An illustration of integrating decision and spatial fusion is shown in Figure 3 with reference window size set to 3×3; the sensor Cmage is preprocessed with Gaussian process resulting in a "Sensor Cmage Patch" similar to the one shown in Figure 1. Note that Gaussian process is not applied to social Cmage, because spontaneous human posts can appear anywhere, making social signals noisy to be modelled with a smoothing kernel as applied for physical sensors readings.

IV. Experiments

A. Dataset

We used the same dataset from our previous work [13], which contains continuous image snapshots from 149 C-CTV traffic cameras across Manhattan, New York City, and geo-tagged tweets posted in a bounding box covers the whole Manhattan area. Each tweet contains the text content, posted time and geographical coordinates (in the form of latitude and longitude). This data set contains dozens of events in various types including protests, festivals, parades, marathons and etc. Different concepts detectors (e.g. parade, people marching, crowd, car etc) are applied to the images, resulting in concept confidence value (ranging from 0 to 1)/ We demonstrate the efficacy of proposed approach on three popular events ("ColumbusDayParade", "MillionMarchNYC" and "StPatricksDayParade") with large spatio-temporal coverage that is examined in work [13].

B. Evaluation Metrics

1) Saliency Metric: Events shown in the image should appear "natural" and "sharp" to a human interpreter [11]. To this point, the fused images are supposed to preserve the salient information and enhance the contrast for visualization. In order to objectively evaluate our hybrid fusion algorithm, we would need a "saliency metric" measure describing how events signals are concentrated in a small dense region. Zhao et al. [17] proposed a feature based metric (Q_P) to measure how well the salient features of source images are preserved. Extending this idea, we define a modified "saliency metric" as a value obtained by averaging the spatial distance of the points belonging to the same cluster with respect to the centroid of the cluster for each cluster. We use mean-shift clustering [3] to obtain clusters from images. Given an image I, *saliency metric* $\mathbf{S}$ is defined by:

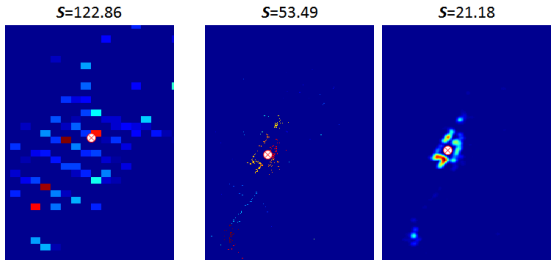
$$\mathbf{S}(I) = \sum_{i=1}^{C} \sum_{\mathbf{p}_m \in CL(c_i)} w_{im} * Dist(\mathbf{p}_m, c_i) \tag{8}$$

where c_i is the cluster centroid of cluster $CL(c_i)$ given by mean-shift clustering and $w_{im} = \frac{e_m}{\sum_{p_n \in CL(c_i)} e_n}$ is the normalized weight for each pixel; for each cluster $\sum w_{im} = 1$.

Lower value of **S** means a more salient and concentrated region, therefore a better image for visual analytics purposes.

2) MSE of Ground Truth: To demonstrate the efficiency of noise removing in fusion process, we evaluate how much the fused Cmage is close to the ground truth manually labelled in our previous work [13].

C. Noise Removal & Saliency Enhancement



(a) sensor cmage (b) social cmage (c) fused result

Figure 4: "Million March NYC" Event Cmages : **(a)** Low Resolution Sensor Cmage of "$Marching$" Concept ; **(b)** High Resolution Social Cmage of "MillionsMarchNYC"; **(c)** Fused Cmage

Fig. 4 shows the sensor(a), social(b) and fused(c) image respectively for the "Million March NYC" protest event. Figure 4a is obtained by applying the "Marching" concept detector on the CCTV camera recordings, generating a low-resolution sensor Cmage. Fig. 4b is obtained by calculating the TF-IDF weight of term "MillionMarchNYC" according to [13], resulting a high-resolution social Cmage. The word cloud examples for three events are shown in Fig. 5 The intensity of the pixels represent the signal strengths at corresponding locations. Red crosses are the centroids of clusters given by the mean-shift algorithm. Saliency metric values are shown on top of the figures. As can be seen, Fig. 4c effectively enhances the contrast and saliency of event candidates than that of Fig. 4a and Fig. 4b, which look noisy. The fused image tells exactly where this marching event is happening. This demonstrates the proposed Bayesian-based fusion in Sec. III-B2 can help enhance the signal if both sources contribute to the confirmation of events and meanwhile eliminates the noise caused by their disagreement.

Figure 5: Word Clouds for Three Different Events.

D. Semantic Details Mining

The effectiveness of fusion is also demonstrated by extensive experiments with different combinations of sensor concepts and social terms (sConcept-term), shown in Fig. 6. Blue, red and green bars are the saliency metric S of sensor, social and fused images respectively. They are ordered by value S of fused images. Rather than presenting only a loosely defined concept, such orderings help users to find the best matching semantic details of ongoing events. For example, details about the "Marching" concept is best described by the social term "blacklivesmatter", which is a popular hashtag posted during the protest. This shows the fusion will have a good performance if two concepts have similar spatial distributions in terms of their event signal.

Table I: ***S*** Values for Different Events

Events	Sensor Cmage	Social Cmage	Fused	Enhancement Rate on Average
ClumbusDayParade	1.24	0.43	0.34	0.47
MillionMarchNYC	1.24	0.47	0.40	0.41
StPatricksDayParade	1.49	0.61	0.53	0.39

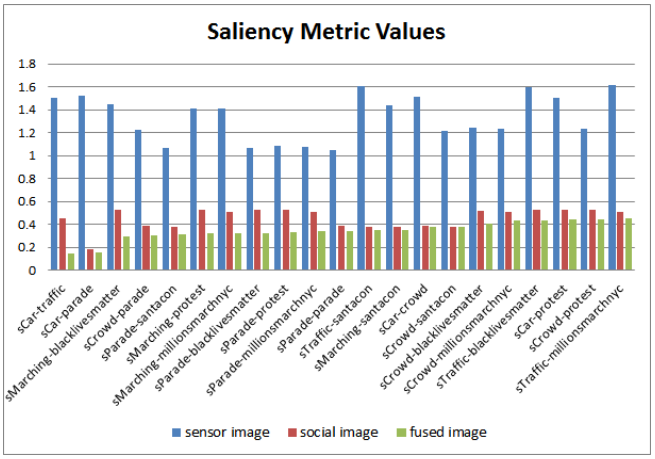


Figure 6: Saliency Metric Values ***S*** of Different Sensor or Social Event Images and Fused Results

Conducting experiments for two more events, we generated Table I showing the average improvement in S values based on the proposed fusion method for different combinations in terms of best matches (e.g. "parade" with "blacklivesmatter", "stpatricksday","green","columbus" etc). The Enhancement Rate measures how much the fused image enhances the saliency for sensor and social on average.

We compare the sensor, social and fused Cmage with ground truth, which is binary picture illustrating the location of this protest event. There are 6 locations where from the camera feeds, we are sure about the event happening and generate a ground truth Cmage accordingly. All Cmages are compared with the ground truth Cmage in terms of MSE. The result is shown in 7.

Since we have detected the events and mined the related semantic words of this situation, we specifically examine the Cmages of concepts that are closely related to this events, including: "crowd","parade","people march-

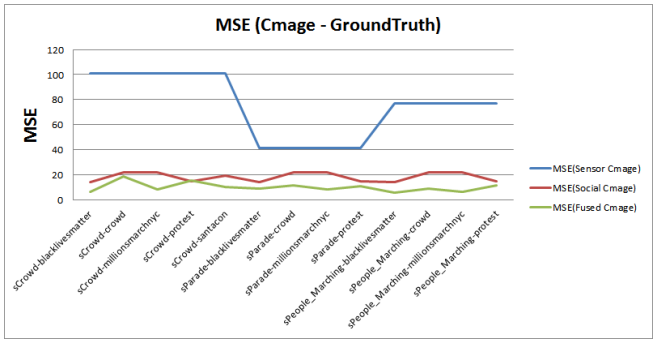


Figure 7: MSE of Sensor, Social and Fused Cmage Compared with Ground Truth Cmage

ing",“blacklivesmatters",“millionmarchnyc" and "protest". As can be seen from Fig. 7, the fused Cmages have less MSE compared to non-fused Cmage, either sensor Cmage of social Cmage. This is because the Bayesian fusion utilizes the agreement the event signals from both sources and the Gaussian process enhance the signal of event locations given their nearby signals contribute to the confirmation of occurring events.

E. Effectiveness of Gaussian process

Sensors sparsity problem is handled by Gaussian process with σ_s^2 set to 0.90 and l_d set to 0.89; these hyper-parameters are learned using the observable sensor Cmage pixels via maximum likelihood estimation [14]. The effectiveness of the Gaussian process for the fusion process is shown in Figure 8, where red line shows the S of fusion without GP and the blue line is the fusion with GP. For the best matches, the fusion will result in better performance (lower S) if Gaussian process is incorporated in the fusion process. However, the fusion of some combinations performs better if no GP is applied. A plausible explanation is that the social term (e.g. santacon) is not semantically related to the sensor concept (describing two different events), so the prediction could not contribute to the fusion.

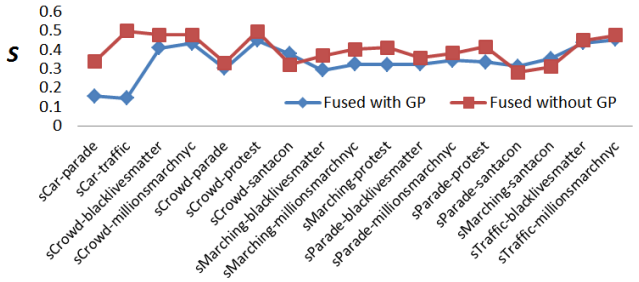


Figure 8: Comparison of Fusion with GP and without GP

V. Conclusions

In this work, we present an image-based hybrid fusion framework to fuse different modalities of physical sensor and social data, considering both the event signal strength and their spatial relations. Image-based representations of different data streams provide not only a better visualization of situations, but also the convenience of manipulation of event-related sensor and social signals. The results demonstrate that the fusion strategy can effectively remove noise from the data streams, localize the event place and offer situational semantics details. For future direction, it would be interesting to explore the semantic relatedness between different social terms and design fusing operators facilitating adding more than two layers.

VI. Acknowledgement

This research was conducted at the NUS-ZJU SeSaMe Centre. It is supported by the Singapore NRF under its IRC@SG Funding Initiative and administered by the IDMPO.

References

[1] P. Atrey, M. Hossain, A. El Saddik, and M. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, 2010.

[2] P. K. Atrey, M. S. Kankanhalli, and R. Jain. Information assimilation framework for event detection in multimedia surveillance systems. *Multimedia systems*, 12(3):239–253, 2006.

[3] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on PAMI*, 24(5):603–619, May 2002.

[4] A. A. Goshtasby and S. Nikolov. Image fusion: Advances in the state of the art. *Information Fusion*, 8(2):114 – 118, 2007.

[5] H.-P. Hsieh, S.-D. Lin, and Y. Zheng. Inferring air quality for station location recommendation based on urban big data. In *KDD*, pages 437–446, 2015.

[6] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. Hauptmann. Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE Transactions on Multimedia*, 12(1):42–53, 2010.

[7] B. Jou, H. Li, J. G. Ellis, D. Morozoff-Abegauz, and S.-F. Chang. Structured exploration of who, what, when, and where in heterogeneous multimedia news sources. In *ACM Multimedia*, pages 357–360, 2013.

[8] Y.-H. Kuo and et al. Discovering the city by mining diverse and multimodal data streams. In *ACM Multimedia*, pages 201–204, 2014.

[9] M. Mazloom, E. Gavves, K. E. A. van de Sande, and C. G. M. Snoek. Searching informative concept banks for video event detection. In *ICMR*, 2013.

[10] B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi. Crowd sensing of traffic anomalies based on human mobility and social media. In *ACM SIGSPATIAL*, pages 344–353, 2013.

[11] G. Piella. Image fusion for enhanced visualization: a variational approach. *IJCV*, 83(1):1–11, 2009.

[12] V. K. Singh, M. Gao, and R. Jain. Social pixels: Genesis and evaluation. In *ACM Multimedia*, pages 481–490, 2010.

[13] Y. Wang and M. S. Kankanhalli. Tweeting cameras for event detection. In *WWW*, pages 1231–1241, 2015.

[14] C. K. Williams and C. E. Rasmussen. Gaussian processes for machine learning. *the MIT Press*, 2(3):4, 2006.

[15] J. Zhang. Multi-source remote sensing data fusion: status and trends. *International Journal of Image and Data Fusion*, 1(1):5–24, 2010.

[16] H. Zhao, Z. Shang, Y. Y. Tang, and B. Fang. Multi-focus image fusion based on the neighbor distance. *Pattern Recognition*, 46(3):1002 – 1011, 2013.

[17] J. Zhao, R. Laganiere, and Z. Liu. Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement. *International Journal of Innovative Computing, Information and Control*, 3(6):1433–1447, 2007.

[18] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian. Query-adaptive late fusion for image search and person re-identification. In *CVPR*, pages 1741–1750, 2015.