Supplementary Materials of Bridging the Ultimate Semantic Gap: A Semantic Search Engine for Internet Videos

Lu Jiang[1], Shoou-I Yu[1], Deyu Meng[2], Teruko Mitamura[1], Alexander G. Hauptmann[1]
[1] School of Computer Science, Carnegie Mellon University
[2] School of Mathematics and Statistics, Xi'an Jiaotong University
{lujiang, iyu, teruko, alex}@cs.cmu.edu, dymeng@mail.xjtu.edu.cn

1. APPENDIX

The supplementary materials provide detailed examples and results in assisting understanding the paper. More information about the feature and the dataset can be found at `http://www.cs.cmu.edu/~lujiang/0Ex/icmr15.html`.

1.1 Full System Configuration

In the multimodal search component, the LM-JM model ($\lambda = 0.7$) is used for ASR/OCR for the frequent-words in the event-kit description. BM25 is used for ASR [8] and OCR features for the event name query (1-3 words), where $k_1 = 1.2$ and $b = 0.75$. Both the frequent-words query and the event name query are automatically generated without manual inspection. While parsing the frequent words in the event-kit description, the stop and template words are first removed, and words in the evidence section are counted three times. After parsing, the words with the frequency ≥ 3 are then used in the query. VSM-tf model is applied to all semantic concept features.

In the SQG component, the exact word matching algorithm finds the concept name in the frequent event-kit words (frequency ≥ 3). The WordNet mapping uses the distance metrics in [10] as the default metric. We build an inverted index over the Wikipedia corpus (about 6 million articles), and use it to calculate the PMI mapping. A pre-trained word embedding trained on Wikipedia [6] is used to calculated the Word embedding mapping.

In the PRF component, the SVM model is selected as the reranking model. The self-paced function used is the mixture weighting:

$$f(\mathbf{v}; k_1, k_2) = -\zeta \sum_{i=1}^{n} \log(v_i + \zeta k_1), \tag{1}$$

where $\zeta = \frac{1}{k_2 - k_1}$ and $k_2 > k_1 > 0$. Its closed-form optimal solution is then written as:

$$v_i^* = \begin{cases} 1 & \ell_i \leq \frac{1}{k_2} \\ 0 & \ell_i \geq \frac{1}{k_1} \\ \frac{\zeta}{\ell_i} - k_1\zeta & \frac{1}{k_2} < \ell_i < \frac{1}{k_1} \end{cases}, \tag{2}$$

where ℓ_i is the loss for the ith sample in the fusion of the ranked lists from multiple modalities. If we rank a list by the sample loss in increasing order, $1/k_1$ is set using the loss of the top $6th$ sample, and $1/k_2$ is set as the loss of the top $3th$ sample. In other words, the top 1-3 samples in the list will have 1.0 weights because their loss $\ell_i \leq 1/k_2$; the top 3-5 samples will have weight $\frac{\zeta}{\ell_i} - k_1\zeta$; the top $6th$ sample will have 0 weights (as its loss $\ell_i = 1/k_1$), and so does those ranked after it. In summary only the top 5 samples have non-zero weights in PRF, and they are used as pseudo positive samples. As discovered in [4] that the pseudo negative samples have neglectable impact on performance. Therefore, we randomly select a number of pseudo negative samples that is proportional to the number of selected pseudo positive samples in each iteration.

When calculating the loss in selecting the pseudo positive samples, it might be the case that more than 5 samples have the loss 0. This happens because the model is not well calibrated. In this case, we tune the parameter b in the loss function so that only one sample has the loss 0. Note that changing interpolation parameter b changes the constant adding to prediction score of each sample, and does not change the rank of the prediction score.

The PRF results by the improved dense trajectory and the MFCC is first averaged. Then the results is averaged with the original ranked list. Since the test set only has around 20 relevant videos in MED14Test, only a single iteration PRF is conducted. For the MED14Eval dataset (200K videos) in our TRECVID 2014 submission, we conduct two iterations of PRF. Generally, we found that the performance starts getting worse after three iterations.

1.2 Query

The input user query is given in the form of the event-kit description by National Institute of Standards and Technology (NIST). Table 5 and Table 4 show the user queries for the event "E011 Making a sandwich" and "E012 Parade". Their corresponding system queries are shown in Table 6 and Table 7. Indeed, as we see, SQG is very challenging because it involves the understanding of the description written in natural language.

We found two interesting observations in the semantic

ICMR '15 Shanghai, China

query generation. First discriminating concept relevance in a query tends to increases the performance. In our system, the relevance is categorized into three levels: "very relevant", "relevant" and "slightly relevant", and are assigned to weights of 2.0, 1.0 and 0.5, respectively. See an example in Table 6 or Table 7. Table 1 compares our full system with and without query weighting. As we see, the system with query weighting outperforms the one without it.

Second, the Boolean logic query tends to improve the performance for certain queries. For example, in the event "E029: Winning a race without a vehicle", the query including only relevant concepts such as swimming, racing or marathon can achieve a MAP of 12.57. However, the query also containing "AND NOT" concepts such as car racing, horse riding or bicycling can achieve a MAP of 24.50. This observation may suggest formulating logic queries seems to be a promising direction, under the circumstance that the concept detector are reasonably accurate.

Table 1: Analysis of weighted and logic queries.

Runs	MED13Test		MED14Test	
	Single	10 Splits	Single	10 Splits
Query Weight	**20.60**	**18.77±2.16**	**20.75**	**19.47±1.19**
NoWeight	18.8	18.30±2.20	20.27	19.35±1.26
E029 Logic	**24.50**	**18.47±3.14**	**24.50**	**18.47±3.14**
E029 NoLogic	12.57	11.89±2.05	12.57	11.89±2.05

1.3 Retrieval System and Results

Fig. 1 shows a screenshot of our prototype system for the query "E012 Parade". The left panel shows the query bucket that contains relevant visual concepts input by the user. The right panel shows the returned videos. As we see, most of the returned videos are relevant to the query.

1.4 Comparison to published methods

To our best knowledge, Table 2 and Table 3 list a comprehensive survey of the published results on zero-example multimedia event detection. These results are all conducted on the NIST's split on MED13Test and MED14Test, and thus are comparable to each other. The last three rows: AutoSQGSys, VisualSys and FullSys are systems proposed in this paper, where AutoSQGSys uses the automatically generated concept mapping; VisualSys uses only visual features; FullSys uses all features but no PRF. The methods are ranked according to their MAPs.

Table 2: MAP (× 100) comparison with the published results on MED13Test.

Method	Year	MAP
SIN/DCNN (visual) [4]	2014	2.5
Composite Concepts [1]	2014	6.4
Tag Propagation [7]	2014	9.6
MMPRF [4]	2014	10.1
Clauses [5]	2014	11.2
Multi-modal Fusion [9]	2014	12.6
SPaR [2]	2014	12.9
Our AutoSQGSys	2015	12.0
Our VisualSys	2015	18.3
Our FullSys	2015	20.7

Table 3: MAP (× 100) comparison with the published results on MED14Test.

Method	Year	MAP
SPCL [3]	2015	9.2
Our AutoSQGSys	2015	11.5
Our VisualSys	2015	17.6
Our FullSys	2015	20.6

1.5 Event-level Contribution

Table 8 lists the event-level modality contribution on the NIST's split. There are 10 overlapping events between the two sets MED13Test and MED14Test. The summary of the MAP can be found at Table 3 in the paper. Table 9 and Table 10 list the event-level contribution about the visual and textual features (APs after being removed from the full system). The summary of the MAP on the two datasets can be found at Table 4 in the paper.

2. REFERENCES

[1] A. Habibian, T. Mensink, and C. G. Snoek. Composite concept discovery for zero-shot video event detection. In *ICMR*, 2014.

[2] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann. Easy samples first: Self-paced reranking for zero-example multimedia search. In *MM*, 2014.

[3] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann. Self-paced curriculum learning. In *AAAI*, 2015.

[4] L. Jiang, T. Mitamura, S.-I. Yu, and A. G. Hauptmann. Zero-example event search using multimodal pseudo relevance feedback. In *ICMR*, 2014.

[5] H. Lee. Analyzing complex events and human actions in" in-the-wild" videos. In *UMD Ph.D Theses and Dissertations*, 2014.

[6] O. Levy and Y. Goldberg. Dependency-based word embeddings. In *ACL*, 2014.

[7] M. Mazloom, X. Li, and C. G. Snoek. Few-example video event retrieval using tag propagation. In *ICMR*, 2014.

[8] Y. Miao, F. Metze, and S. Rawat. Deep maxout networks for low-resource speech recognition. In *ASRU*, 2013.

[9] S. Wu, S. Bondugula, F. Luisier, X. Zhuang, and P. Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *CVPR*, 2014.

[10] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *ACL*, 1994.

[11] S. Xu, H. Li, X. Chang, S.-I. Yu, X. Du, X. Li, L. Jiang, Z. Mao, Z. Lan, S. Burger, and A. Hauptmann. Incremental multimodal query construction for video search. In *ICMR*, 2015.

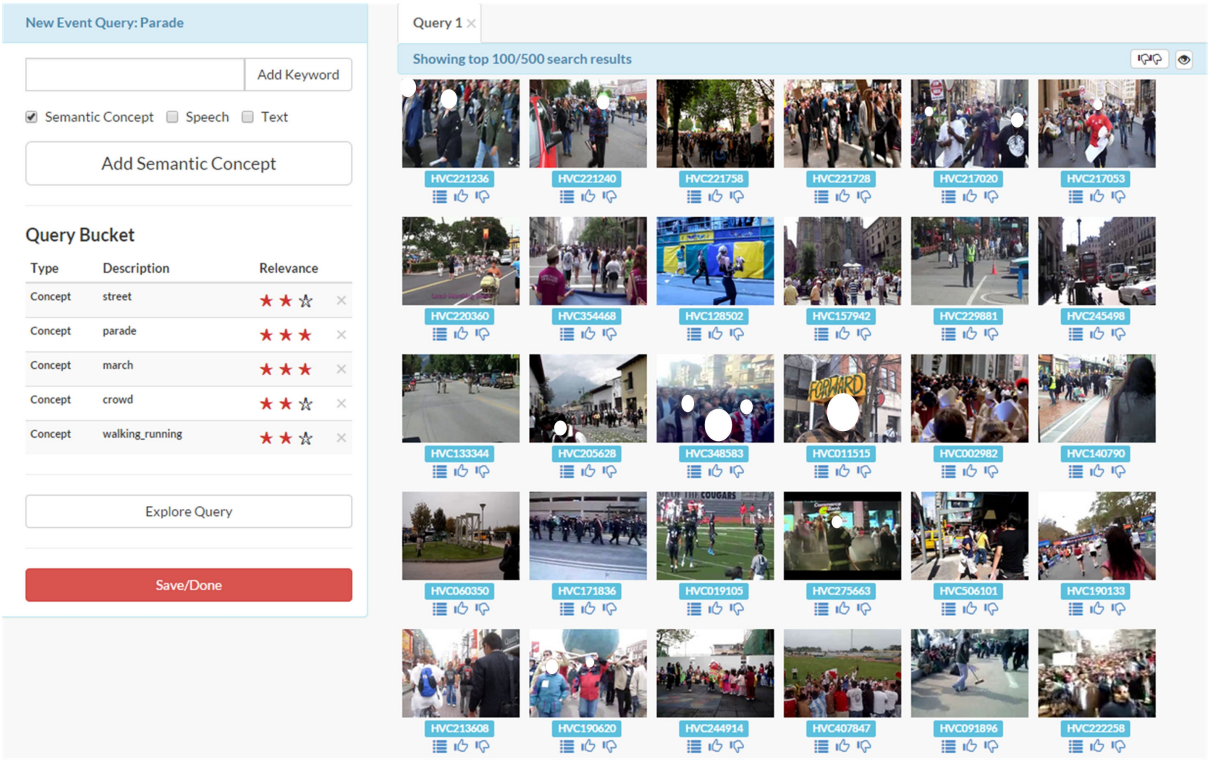

Figure 1: The screenshot of our E-Lamp 0Ex system for the query "E012 Parade" [11]. The left panel shows the query bucket that contains relevant visual concepts input by the user. The right panel shows the returned results.

Table 4: User query (event-kit description) for the event "E012 Parade".

Event name		Parade
Definition		A large group of people process for a celebration/commemorationof some event
Explication		A parade is a group of people processing either in celebration or commemoration of some event. Parades typically involve one or more groups of people proceeding down a route between two lines of spectators. Most often parades process down a street and the spectators are lined up on either side of the street. People in the parade may be driving cars, riding horses, and/or walking, dancing as a group in coordinated special dress or costumes, or riding on a parade float. Parade floats are decorated platforms that sit on top of a vehicle or are pulled by a vehicle or by people as part the procession. Parades are generally accompanied by music and by cheering or clapping from the spectators. Military groups may participate in parades, but not all military demonstrations constitute a parade.
Evidences	scene	typically outdoors, any season, usually on a street
	objects/people	a very large group of people, with floats, costumes, props, vehicles, horses, megaphones
	activities	marching, walking, singing, dancing, clapping, yelling
	audio	music from bands; crowd cheering or clapping; announcers describing the goings-on; horns or other vehicle noises

Table 5: User query (event-kit description) for the event "Making a sandwich".

Event name		Making a sandwich
Definition		Constructing an edible food item from ingredients, often including one or more slices of bread plus fillings
Explication		Sandwiches are generally made by placing food items on top of a piece of bread, roll or similar item, and placing another piece of bread on top of the food items. Sandwiches with only one slice of bread are less common and are called "open face sandwiches". The food items inserted within the slices of bread are known as "fillings" and often include sliced meat, vegetables (commonly used vegetables include lettuce, tomatoes, onions, bell peppers, bean sprouts, cucumbers, and olives), and sliced or grated cheese. Often, a liquid or semi-liquid "condiment" or "spread" such as oil, mayonnaise, mustard, and/or flavored sauce, is drizzled onto the sandwich or spread with a knife on the bread or top of the sandwich fillers. The sandwich or bread used in the sandwich may also be heated in some way by placing it in a toaster, oven, frying pan, countertop grilling machine, microwave or grill. Sandwiches are a popular meal to make at home and are available for purchase in many cafes, convenience stores, and as part of the lunch menu at many restaurants.
Evidences	scene	indoors (kitchen or restaurant or cafeteria) or outdoors (a park or backyard)
	objects/people	bread of various types; fillings (meat, cheese, vegetables), condiments, knives, plates, other utensils
	activities	slicing, toasting bread, spreading condiments on bread, placing fillings on bread, cutting or dishing up fillings
	audio	noises from equipment hitting the work surface; narration of or commentary on the process; noises emanating from equipment (e.g. microwave or griddle)

Table 6: System query for the event "E011 Making a sandwich".

Event	ID	Name	Category	Relevance
Visual	sin346_133	food	man made thing, food	very relevant
	sin346_183	kitchen	structure building, room	very relevant
	yfcc609_505	cooking	human activity, working utensil tool	very relevant
	sin346_261	room	structure building, room	relevant
	sin346_28	bar_pub	structure building,commercial building	relevant
	yfcc609_145	lunch	food, meal	relevant
	yfcc609_92	dinner	food, meal	relevant
ASR	ASR_long	sandwich, food, bread, fill, place, meat, vegetable, cheese, condiment, knife, plate, utensil, slice, toast, spread, cut, dish	-	relevant
OCR	OCR_short	sandwich	-	relevant

Table 7: System query for the event "E012 Parade". ASR/OCR is not used in this event.

Event	ID	Name	Category	Relevance
Visual	sin346_299	street	structure building,transport structure	very relevant
	sin346_229	people_marching	human activity, marching	very relevant
	sin346_83	crowd	human-features, number of people	very relevant
	yfcc609_34	parade	human activity, parade rally	very relevant
	yfcc644_213	parade	human activity, parade rally	very relevant
	sin346_130	flag	man made thing, textile object	relevant
	sin346_296	standing	human activity	slightly relevant
	sin346_338	walking	human activity, walking sports	slightly relevant
ASR	-	-	-	-
OCR	-	-	-	-

Table 8: Event-level comparison of modality contribution on the NIST split. The best AP is marked in bold.

Event ID & Name	FullSys	FullSys+PRF	VisualSys	ASRSys	OCRSys
E006: Birthday party	0.3842	**0.3862**	0.3673	0.0327	0.0386
E007: Changing a vehicle tire	0.2322	**0.3240**	0.2162	0.1707	0.0212
E008: Flash mob gathering	0.2864	**0.4310**	0.2864	0.0052	0.0409
E009: Getting a vehicle unstuck	**0.1588**	0.1561	**0.1588**	0.0063	0.0162
E010: Grooming an animal	**0.0782**	0.0725	**0.0782**	0.0166	0.0050
E011: Making a sandwich	0.1183	0.1304	0.1064	**0.2184**	0.0682
E012: Parade	**0.5566**	0.5319	**0.5566**	0.0080	0.0645
E013: Parkour	0.0545	**0.0839**	0.0448	0.0043	0.0066
E014: Repairing an appliance	0.2619	**0.2989**	0.2341	0.2086	0.0258
E015: Working on a sewing project	**0.2068**	0.2021	**0.2036**	0.0866	0.0166
E021: Attempting a bike trick	0.0635	**0.0701**	0.0635	0.0006	0.0046
E022: Cleaning an appliance	**0.2634**	0.1747	0.0008	**0.2634**	0.0105
E023: Dog show	**0.6737**	0.6610	**0.6737**	0.0009	0.0303
E024: Giving directions to a location	**0.0614**	0.0228	0.0011	**0.0614**	0.0036
E025: Marriage proposal	0.0188	**0.0270**	0.0024	0.0021	0.0188
E026: Renovating a home	**0.0252**	0.0160	**0.0252**	0.0026	0.0023
E027: Rock climbing	**0.2077**	0.2001	0.2077	0.1127	0.0038
E028: Town hall meeting	0.2492	**0.3172**	0.2492	0.0064	0.0134
E029: Winning a race without a vehicle	0.1257	**0.1929**	0.1257	0.0011	0.0019
E030: Working on a metal crafts project	0.1238	**0.1255**	0.0608	0.0981	0.0142
E031: Beekeeping	0.5883	**0.6401**	0.5883	0.2676	0.0440
E032: Wedding shower	0.0833	**0.0879**	0.0459	0.0428	0.0017
E033: Non-motorized vehicle repair	0.5198	**0.5263**	0.5198	0.0828	0.0159
E034: Fixing musical instrument	0.0276	**0.0444**	0.0170	0.0248	0.0023
E035: Horse riding competition	0.3677	**0.3710**	0.3677	0.0013	0.0104
E036: Felling a tree	0.0968	**0.1180**	0.0968	0.0020	0.0076
E037: Parking a vehicle	**0.2918**	0.2477	**0.2918**	0.0008	0.0009
E038: Playing fetch	0.0339	**0.0373**	0.0339	0.0020	0.0014
E039: Tailgating	0.1437	**0.1501**	0.1437	0.0013	0.0388
E040: Tuning musical instrument	0.1554	**0.3804**	0.0010	0.1840	0.0677
MAP (MED13Test E006-E015 E021-E030)	0.2075	**0.2212**	0.1831	0.0653	0.0203
MAP (MED14Test E021-E040)	0.2060	**0.2205**	0.1758	0.0579	0.0147

Table 9: Event-level comparison of visual feature contribution on the NIST split.

Event ID & Name	FullSys	MED/IACC	MED/Sports	MED/YFCC	MED/DIY	MED/ImageNet
E006: Birthday party	0.3842	0.3797	0.3842	0.2814	0.3842	0.2876
E007: Changing a vehicle tire	0.2322	0.2720	0.2782	0.1811	0.1247	0.0998
E008: Flash mob gathering	0.2864	0.1872	0.2864	0.3345	0.2864	0.2864
E009: Getting a vehicle unstuck	0.1588	0.1070	0.1588	0.1132	0.1588	0.1588
E010: Grooming an animal	0.0782	0.0902	0.0782	0.0914	0.0474	0.0782
E011: Making a sandwich	0.1183	0.0926	0.1183	0.1146	0.1183	0.1183
E012: Parade	0.5566	0.5738	0.5566	0.3007	0.5566	0.5566
E013: Parkour	0.0545	0.0066	0.0545	0.0545	0.0545	0.0545
E014: Repairing an appliance	0.2619	0.2247	0.2619	0.1709	0.2619	0.1129
E015: Working on a sewing project	0.2068	0.2166	0.2068	0.2068	0.1847	0.0712
E021: Attempting a bike trick	0.0635	0.0635	0.0006	0.0635	0.0635	0.0635
E022: Cleaning an appliance	0.2634	0.2634	0.2634	0.2634	0.2634	0.2634
E023: Dog show	0.6737	0.6737	0.0007	0.6737	0.6737	0.6737
E024: Giving directions to a location	0.0614	0.0614	0.0614	0.0614	0.0614	0.0614
E025: Marriage proposal	0.0188	0.0188	0.0188	0.0188	0.0188	0.0188
E026: Renovating a home	0.0252	0.0017	0.0252	0.0252	0.0252	0.0252
E027: Rock climbing	0.2077	0.2077	0.0009	0.2077	0.2077	0.2077
E028: Town hall meeting	0.2492	0.0956	0.2492	0.2418	0.2492	0.2492
E029: Winning a race without a vehicle	0.1257	0.1257	0.0056	0.1257	0.1257	0.1257
E030: Working on a metal crafts project	0.1238	0.1238	0.1238	0.0981	0.1238	0.1238
E031: Beekeeping	0.5883	0.5883	0.5883	0.5883	0.5883	0.0012
E032: Wedding shower	0.0833	0.0833	0.0833	0.0833	0.0924	0.0833
E033: Non-motorized vehicle repair	0.5198	0.5198	0.4440	0.5198	0.4742	0.4417
E034: Fixing musical instrument	0.0276	0.0276	0.0276	0.0276	0.0439	0.0276
E035: Horse riding competition	0.3677	0.3430	0.1916	0.3677	0.3677	0.3677
E036: Felling a tree	0.0968	0.0275	0.1100	0.0968	0.0968	0.0968
E037: Parking a vehicle	0.2918	0.1902	0.2918	0.2918	0.2918	0.1097
E038: Playing fetch	0.0339	0.0339	0.0008	0.0339	0.0339	0.0339
E039: Tailgating	0.1437	0.0631	0.1437	0.0666	0.1437	0.1437
E040: Tuning musical instrument	0.1554	0.1554	0.1554	0.1554	0.1554	0.1554
MAP (MED13Test E006-E015 E021-E030)	0.2075	0.1893	0.1567	0.1814	0.1995	0.1818
MAP (MED14Test E021-E040)	0.2060	0.1834	0.1393	0.2005	0.2050	0.1637

Table 10: Event-level comparison of textual feature contribution on the NIST split.

Event ID & Name	FullSys	MED/ASR	MED/MED
E006: Birthday party	0.3842	0.3842	0.3673
E007: Changing a vehicle tire	0.2322	0.2162	0.2322
E008: Flash mob gathering	0.2864	0.2864	0.2864
E009: Getting a vehicle unstuck	0.1588	0.1588	0.1588
E010: Grooming an animal	0.0782	0.0782	0.0782
E011: Making a sandwich	0.1183	0.1043	0.1205
E012: Parade	0.5566	0.5566	0.5566
E013: Parkour	0.0545	0.0545	0.0448
E014: Repairing an appliance	0.2619	0.2436	0.2527
E015: Working on a sewing project	0.2068	0.1872	0.2242
E021: Attempting a bike trick	0.0635	0.0635	0.0635
E022: Cleaning an appliance	0.2634	0.0008	0.2634
E023: Dog show	0.6737	0.6737	0.6737
E024: Giving directions to a location	0.0614	0.0011	0.0614
E025: Marriage proposal	0.0188	0.0188	0.0024
E026: Renovating a home	0.0252	0.0252	0.0252
E027: Rock climbing	0.2077	0.2077	0.2077
E028: Town hall meeting	0.2492	0.2492	0.2492
E029: Winning a race without a vehicle	0.1257	0.1257	0.1257
E030: Working on a metal crafts project	0.1238	0.0608	0.1238
E031: Beekeeping	0.5883	0.5883	0.5883
E032: Wedding shower	0.0833	0.0833	0.0459
E033: Non-motorized vehicle repair	0.5198	0.5198	0.5198
E034: Fixing musical instrument	0.0276	0.0314	0.0178
E035: Horse riding competition	0.3677	0.3677	0.3677
E036: Felling a tree	0.0968	0.0968	0.0968
E037: Parking a vehicle	0.2918	0.2918	0.2918
E038: Playing fetch	0.0339	0.0339	0.0339
E039: Tailgating	0.1437	0.1437	0.1437
E040: Tuning musical instrument	0.1554	0.0893	0.1840
MAP (MED13Test E006-E015 E021-E030)	0.2075	0.1848	0.2059
MAP (MED14Test E021-E040)	0.2060	0.1836	0.2043