

Multimedia Event Detection: Strong by Integration

Hao ZHANG¹, Maaïke de Boer²
Yijie Lu¹, Klamer Schutte², Wessel Kraaij², Chong-Wah Ngo¹

¹City University of Hong Kong

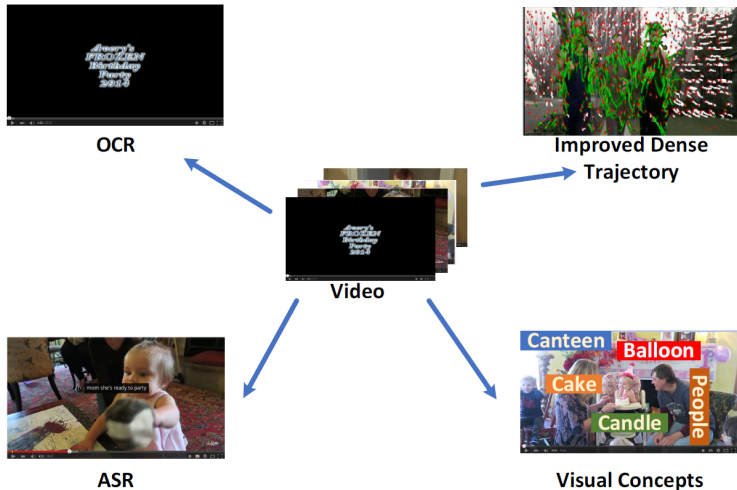
²TNO and Radboud University

November 24, 2015

- Observations
- Modalities
- System
- Fusion: Joint Probability
- Fusion: Adding Zero-Shot
- Reranking: OCR/ASR
- Experiments: MED14_Test/MED15_Eval
- Conclusion

Observations

As is well known, multimedia event consists of **multi-modalities**:
Audio, Motion, Visual, Texts ...



Multi-modalities: Audio, Motion, Visual, Texts ...

Multi-modalities: Audio, Motion, Visual, Texts ...

More efforts: **single**-modality.

Multi-modalities: Audio, Motion, Visual, Texts ...

More efforts: **single**-modality. e.g:

- Motion features: Dense Trajectories, Improved Dense Trajectories.
- Visual features: HOG, SIFT, Deep Features ...

Multi-modalities: Audio, Motion, Visual, Texts ...

More efforts: **single**-modality. e.g:

- Motion features: Dense Trajectories, Improved Dense Trajectories.
- Visual features: HOG, SIFT, Deep Features ...

Less efforts: **integrate** across modalities.

Problem:

Intergrating across modalities

Problem:

Intergrating across modalities

Difficulties:

- Modalities have different meanings.
- Modalities have different precisions.

Modalities

OCR



ASR



Concepts



IDT



Balloon
Canteen
Candle
People
Cake

1. Semantics
2. High Precision
3. Relevant
4. Low Recall (rare)



Modalities

OCR



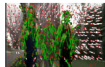
ASR



Concepts



IDT



1. Semantics
2. High Precision (wanted)
3. Relevant/Irrelevant
4. Low Recall (rare)



Modalities

OCR



ASR



Concepts



IDT



1. Semantics
2. Low Precision
3. Low Recall
4. Relevant/Irrelevant
5. High Recall (many)



Modalities

OCR



ASR



Concepts



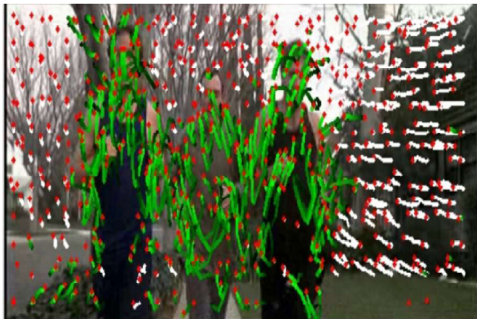
IDT



Balloon
Canteen
Candle
People
Cake



1. Non-Semantics
2. Low precision
3. Relevant/Irrelevant
4. High Recall (many)



Modalities

OCR



ASR

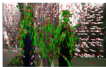


Concepts



Balloon
Canteen
Candle
People

IDT



1. Semantics
2. High Precision
3. Relevant
4. Low Recall

Reranking

1. Semantics
2. High Precision
3. Relevant/Irrelevant
4. Low Recall

1. Semantics
2. Low Precision
3. Relevant/Irrelevant
4. High Recall

Fusion

1. Non-Semantics
2. Low Precision
3. Relevant/irrelevant
4. High Recall

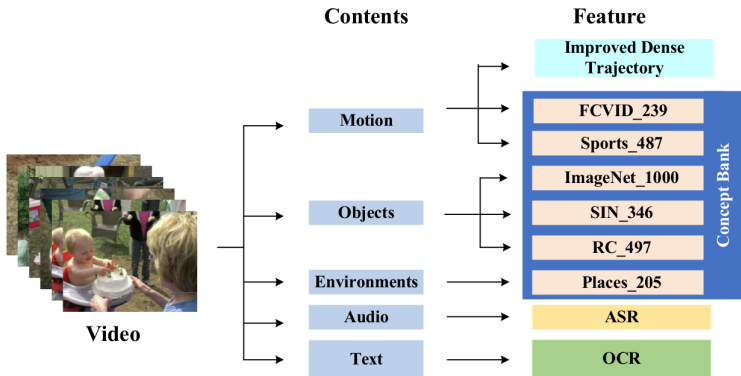
For Event Detection with 100Ex/10Ex:

An **intergration system with multi-modalities**.

We present 100Ex/10Ex as:

- Multi-modalities
- Different methods for different modalities
- Integration of modalities

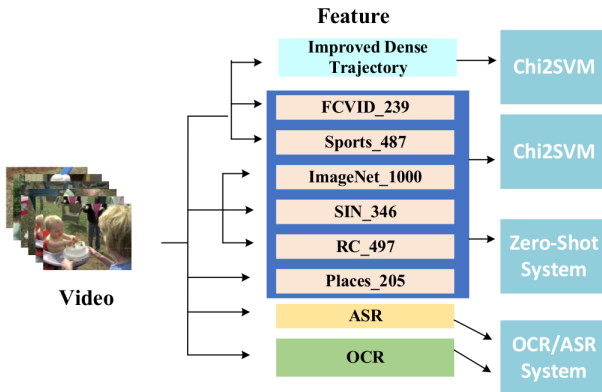
Modalities



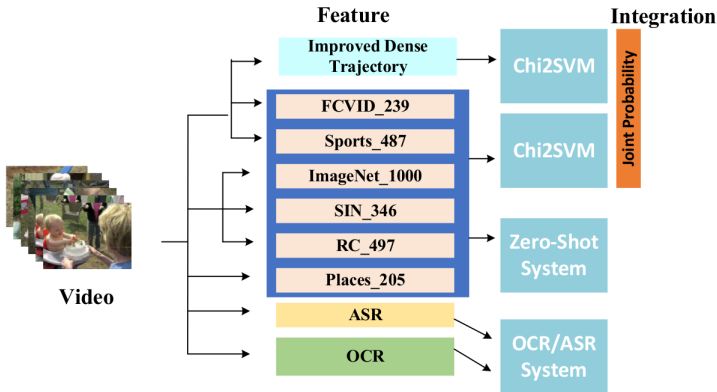
Concept Bank

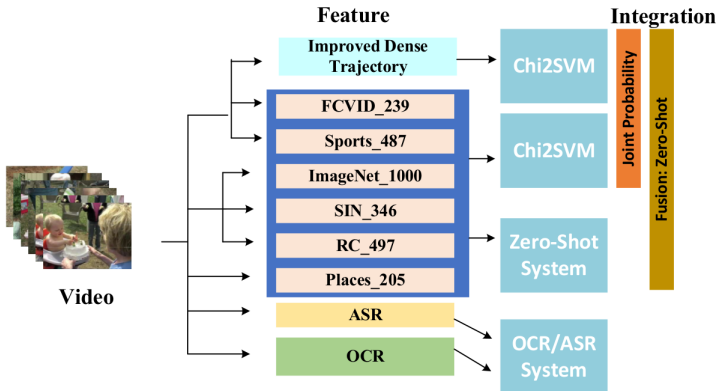
| Feature | Dim | Structure | Dataset |
|---------------|------|-----------|------------------------|
| Sports_487 | 487 | 3D-CNN | Sports-1M |
| ImageNet_1000 | 1000 | DCNN | ImageNet |
| SIN_346 | 346 | DCNN | TRECVID SIN |
| RC_487 | 487 | DCNN | TRECVID Research Set |
| Places_205 | 205 | DCNN | MIT Places |
| FCVID_239 | 239 | SVM | Fudan-Columbia Dataset |

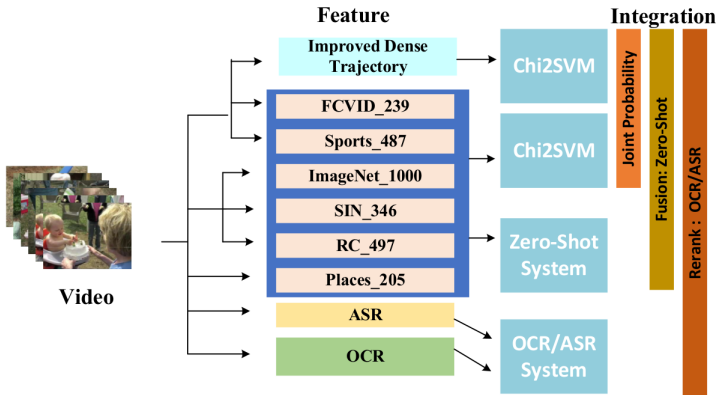
Table : Concept Bank



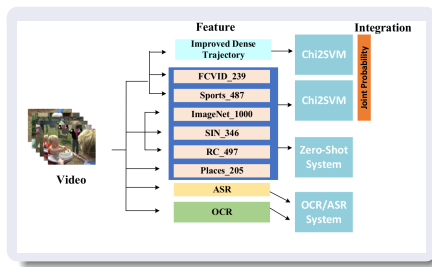
We propose three stages of fusion strategy, which can improve event detection step-by-step.







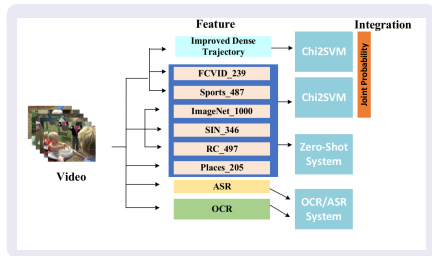
Fusion: Joint Probability



Classification:

Two classifiers make predicts independently.

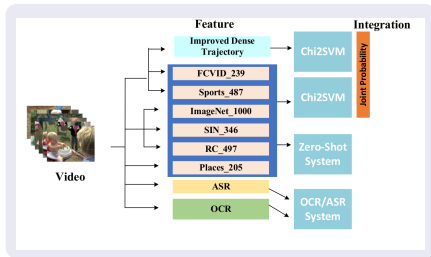
Fusion: Joint Probability



Average:

A low score of one type of classifier downgrades a possibly relevant video.

Fusion: Joint Probability



Average:

A low score of one type of classifier downgrades a possibly relevant video.

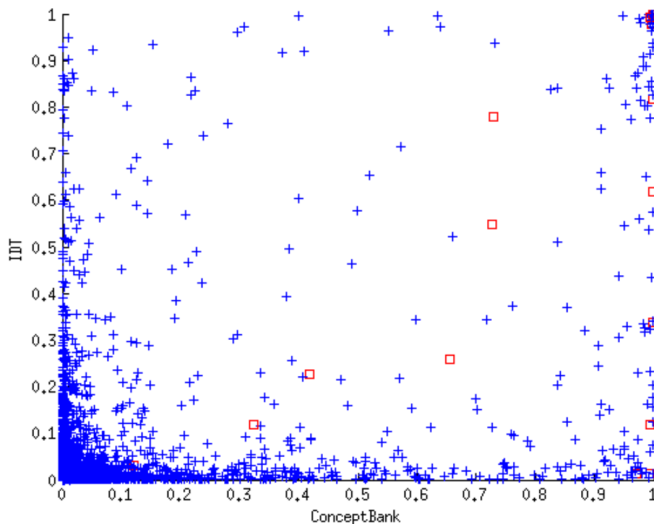
Joint Probability:

Only videos that receive a low score from both classifiers will be put at the bottom of the ranking list.

$$JP = 1 - (1 - P_{CB}) \times (1 - P_{IDT})$$

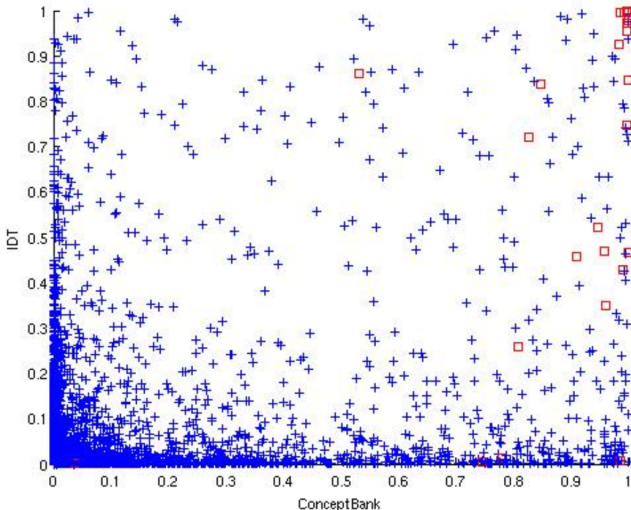
Fusion:Joint Probability

E021-SVM Prediction Scores with Concept feature and Improved Dense Trajectory



Fusion: Joint Probability

E039-SVM Prediction Scores with Concept feature and Improved Dense Trajectory

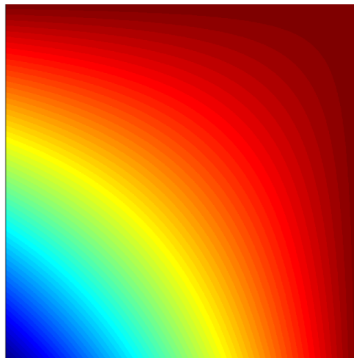


Contour Map

Average Fusion

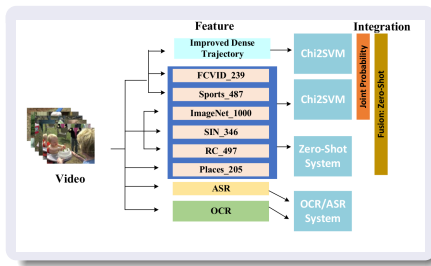


Joint Probability



- Joint Probability is our first try to fuse two kinds of prediction scores by distributions of predicted scores.
- Based on the distributions of predicted scores, there might be more powerful unsupervised distribution-based fusion strategy.

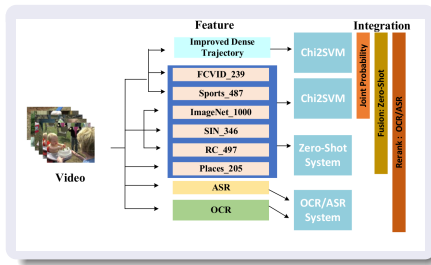
Fusion: Adding Zero-Shot



Adding Zero-Shot:

We averaged scores predicted by the Zero-Shot system (the other PPT) with scores predicted by the event detectors (SVM).

Reranking: OCR/ASR



"Re-ranking":

Design **high precision** ASR and OCR systems for reranking.

Reranking: OCR/ASR

Recall OCR

OCR



ASR



Concepts



IDT



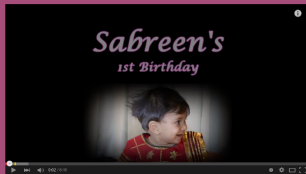
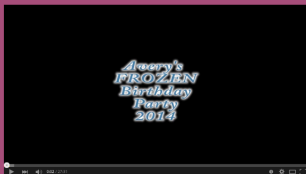
Balloon
Canteen
Candle
Cake
People

1. Semantics
2. High Precision
3. Relevant
4. Low Recall (rare)



OCR Observations:

Birthday Party



Bee Keeping



OCR Observations:

Tuning Musical Instruments



Bike Trick



Rock Climbing



OCR Observations and Strategy:

- Parts of relevant videos were **post-produced** (include **titles**).

OCR Observations and Strategy:

- Parts of relevant videos were **post-produced** (include **titles**).
- Pick out these video by matching OCR and Query.

OCR Observations and Strategy:

- Parts of relevant videos were **post-produced** (include **titles**).
- Pick out these video by matching OCR and Query.
- Rerank these videos with extra-bonus score, boosting their ranks.

OCR Observations and Strategy:

- Parts of relevant videos were **post-produced** (include **titles**).
- Pick out these video by matching OCR and Query.
- Rerank these videos with extra-bonus score, boosting their ranks.

Same strategy is used for ASR,

Reranking: OCR/ASR

Drawbacks of ASR:

OCR



ASR



Concepts



IDT

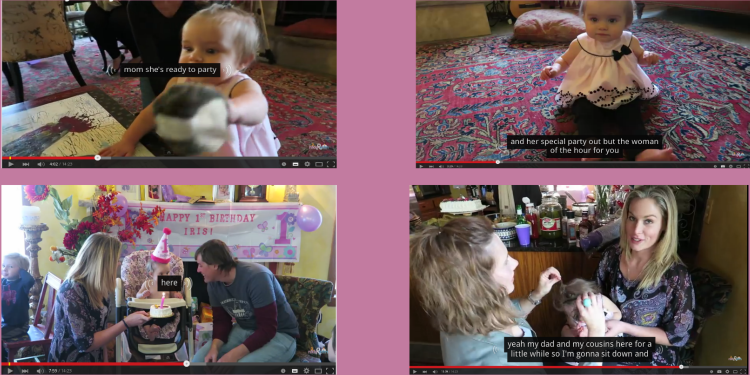


1. Semantics
2. High Precision (wanted)
3. Relevant/Irrelevant
4. Low Recall (rare)



ASR Observations:

Birthday Party



The figure displays four video frames from a birthday party, each with an ASR subtitle overlaid. The frames are arranged in a 2x2 grid. The top-left frame shows a baby sitting on a table, with the subtitle "mom she's ready to party". The top-right frame shows a baby standing on a patterned rug, with the subtitle "and her special party out but the woman of the hour for you". The bottom-left frame shows a woman sitting at a table with a birthday cake, with the subtitle "here". The bottom-right frame shows a woman holding a baby, with the subtitle "yeah my dad and my cousins here for a little while so I'm gonna sit down and".

ASR Observations:

Bike Trick



VIREO Video Retrieval Group

TNO innovation
for life

Radboud Universiteit Nijmegen



ASR Observations:

- The portion of **relevant** ASR results is **small**.

ASR Observations:

- The portion of **relevant** ASR results is **small**.
- The portion of **irrelevant** ASR results is **large**.

ASR Observations:

- The portion of **relevant** ASR results is **small**.
- The portion of **irrelevant** ASR results is **large**.
- Mining event relevance with ASR is still an open topic.

The indexing and search tool Lucene is used for the OCR and ASR data. High precision is retrieved by:

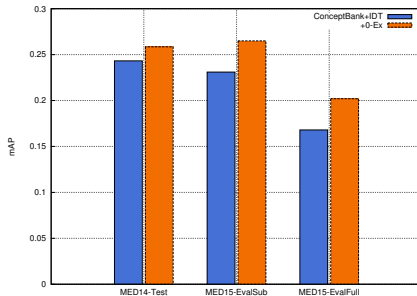
- OCR: manually defining a Boolean Query using the event description and Wikipedia and some information on known common mistakes from the Tesseract tool (e.g. zero (0) and O).
- ASR: manually defining a Boolean Query and adding a PhraseQuery so the words in the query do not occur more than five words from each other. Only the words specific for the event are added.

Based on internal test, we have the following settings for MED 2015 Submission:

- 10 Exemplars:
Adding Zero-Shot, Reranking by OCR/ASR
- 100 Exemplars:
Joint Probability, Reranking by OCR/ASR

Experiments: MED14_Test/MED15_Eval

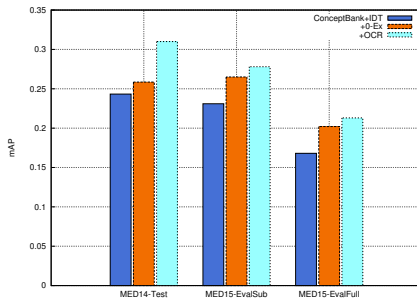
MED PS_10-Ex: Mean AP of fusion strategies on MED14-Test/EvalSub/Full



For 10 exemplars, adding the results of Zero-Shot case does really improve performance (more than 3%)

Experiments: MED14_Test/MED15_Eval

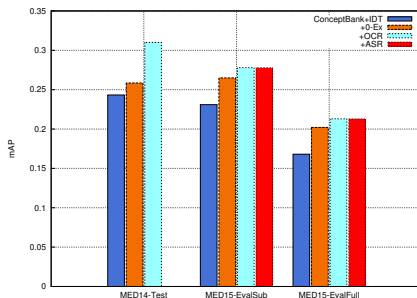
MED PS_10-Ex: Mean AP of fusion strategies on MED14-Test/EvalSub/Full



OCR gives an improvement of 1.2%.

Experiments: MED14_Test/MED15_Eval

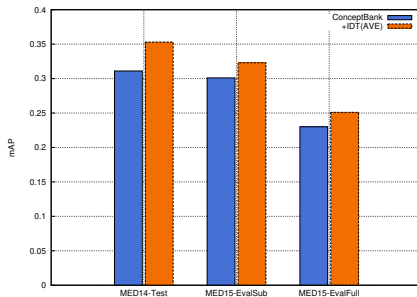
MED PS_10-Ex: Mean AP of fusion strategies on MED14-Test/EvalSub/Full



ASR slightly decreases performance in the Evaluation Set. This is probably because the precision of our ASR system is not as high as our OCR system.

Experiments: MED14_Test/MED15_Eval

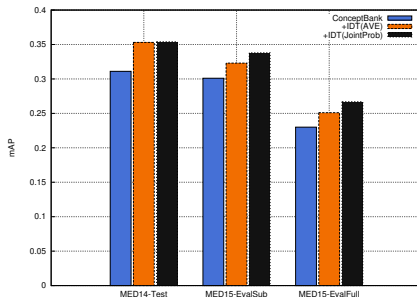
MED PS_100-Ex: Mean AP of fusion strategies on MED14-Test/EvalSub/Full



IDT increases overall performance (2%-4%)

Experiments: MED14_Test/MED15_Eval

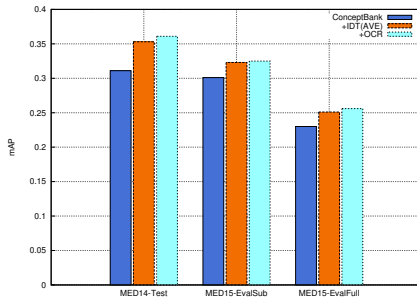
MED PS_100-Ex: Mean AP of fusion strategies on MED14-Test/EvalSub/Full



Joint Probability is better than average fusion, providing for an additional improvement of 1%.

Experiments: MED14_Test/MED15_Eval

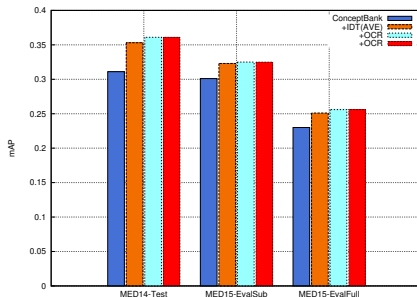
MED PS_100-Ex: Mean AP of fusion strategies on MED14-Test/EvalSub/Full



Adding OCR gives a small improvement.

Experiments: MED14_Test/MED15_Eval

MED PS_100-Ex: Mean AP of fusion strategies on MED14-Test/EvalSub/Full



ASR slightly decreases performance as with the 10Ex Experiments.

- For the 10 Ex case, fusion of the system trained on 10 examples and the zero-shot case improves a lot.
- Fusion with OCR slightly improves performance in all runs. Because the precision of ASR system is not as high as OCR system, performance drops a bit by adding ASR.
- Improved Dense Trajectory improves performance, especially with more training data (100 Ex VS 10 Ex).
- Using Joint Probability of concept features and IDT improves performance on 100 Ex task.