Multimodal Detection, Retrieval and Classification of Social Events in Web Photo Collection

M. Brenner, E. Izquierdo

Multimedia and Vision Research Group
Queen Mary University of London, UK

ICMR 2014 SEWM

Objective

Identify and **retrieve photos** …

… in Collaborative **Web Photo Collections** …

… that are associated with **social events** …

… by exploiting contextual cues and constraints of events …

… and **understand** to which event types they adhere

Introduction and background

- Internet enables people to host, access and share their photos online; for example, through websites like Flickr and Facebook → photos linked to their users
- Collaborative annotations and tags as well as public comments are commonplace, but usually uncontrolled
- Information people assign varies greatly but often seems to include some sort of references to *what* happened *where* and *who* was involved

  → observed experiences or occurrences
  → simply referred to as social events

Introduction and background

Benefits and use-cases of event-driven approaches:

- Easier to search through photo collections if photos are grouped into events
- Possible to link photos/events in web photo collections to public social media like online news feeds
- Reverse: automatically online link news with shared photos

Social events

- Primarily target social events that are public and attended by many people (likely to be better represented in online social media)
- Do not pay attention to personal events (i.e. private vacation trips of individuals)

Social events

The foremost domains defining a social event are:

- Date and time
- Venue (geographic location)
- Involved people …
- … and their observable activities

Observation and assumption

Date and time of capture?

- Most devices store it using EXIF metadata
- Typically embedded in photos

Observation and assumption

Venue/geographic location?

- Like data and time, smartphones embed the geographic location nowadays
- Not always embedded (e.g. if unable to fix GPS)
- Most regular cameras do not store the location

→ only partially available

Observation and assumption

Involved people?

- Analysing photos to determine which people are depicted (face recognition) and thus involved in a social event is difficult, especially when people are not known beforehand

- However: valid assumption that users who upload and share photos are, or were, people involved

- Collaborative photo services like Flickr use unique identifiers (usernames) for their users
  → able to associate each photo with an user

Observation and assumption

Observed activities?

- Captured by photographers
- Implied by visual content of photos
- Implied by collaborative annotations (etc., tags, …)

Event Detection and Retrieval

Define an event as a distinct combination of a spatial window and a temporal window

Basic approach: assume a spatio-temporal cluster is an event

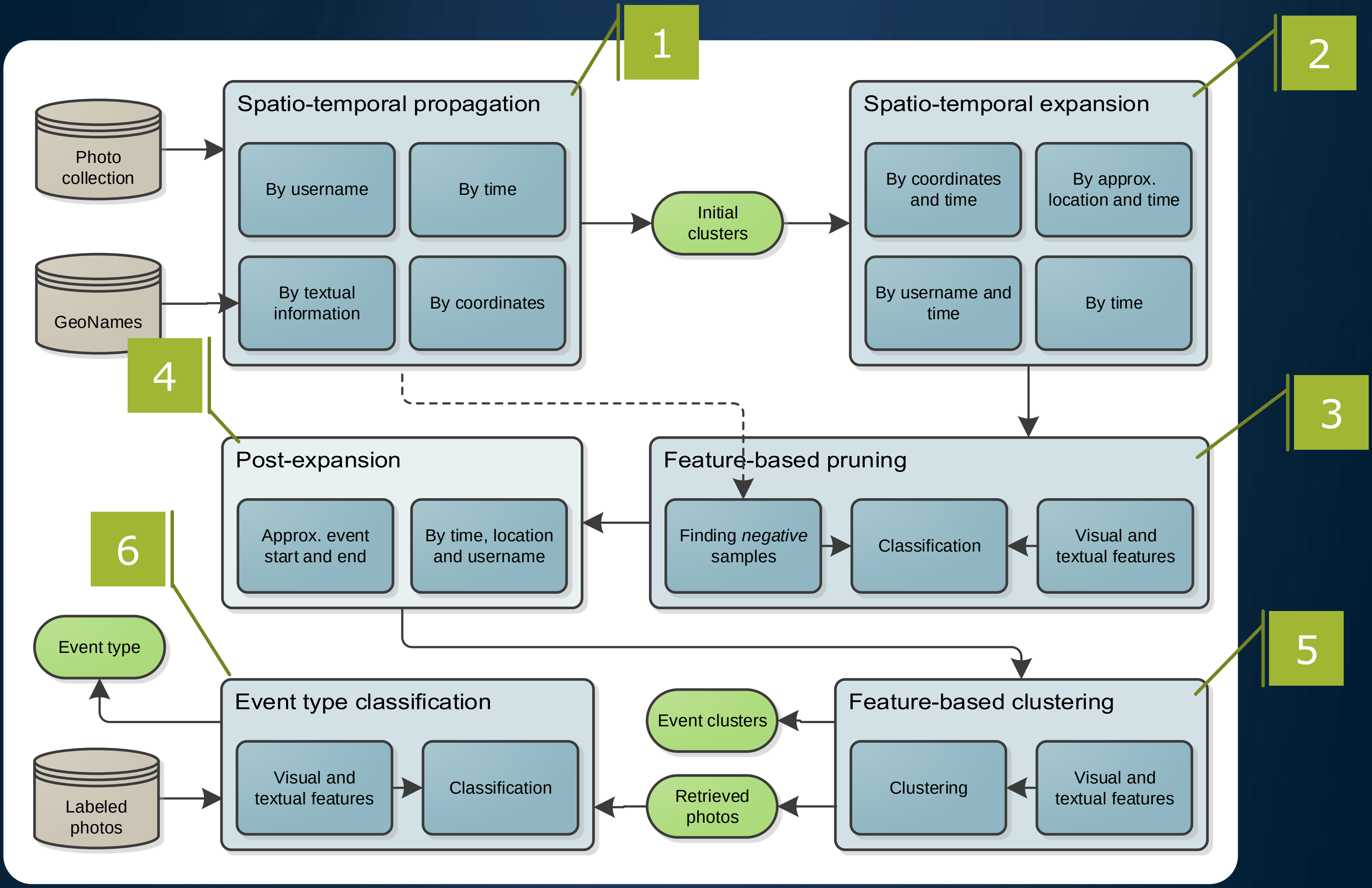
Problem: limited to photos that embed time and location

Extension I: extend to remaining photos not including location

- extend spatio-temporal clusters → retrieval space
- Select or prune non-related photos by feature-based classification → model topic (observed activities)
- Re-include mistakenly discarded photos based on usernames/time

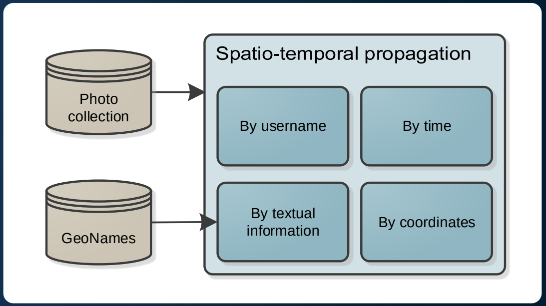
Extension II: additional clustering

Overview of framework



Spatio-temporal propagation

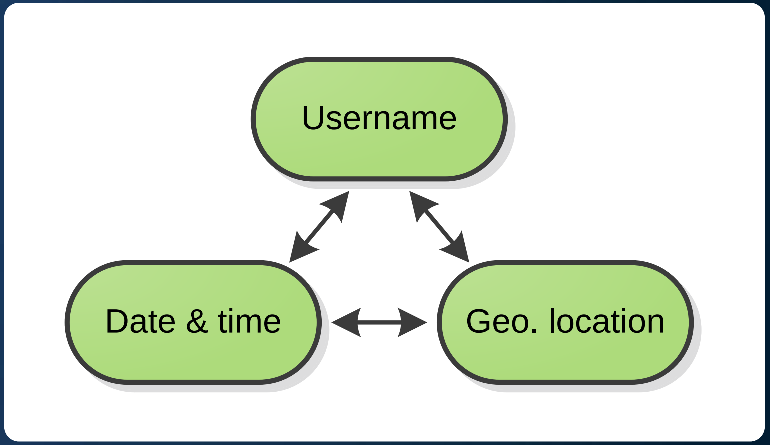
- Most traditional cameras still lack capability of determining the geographic location
- Smartphones usually offer this capability, but cannot provide location information at all times (e.g. GPS signals within buildings are often too weak to fix the location)
- to still determine the location of as many photos as possible → propagate location from photos that embed location to those that do not



Assumption

Due to contextual constraints:

- Photos sharing the same username, date and time as well as geographical location shall belong together to the same event
- Likewise, photos that differ in at least one constraint shall not belong together



Propagating location: *Exact*

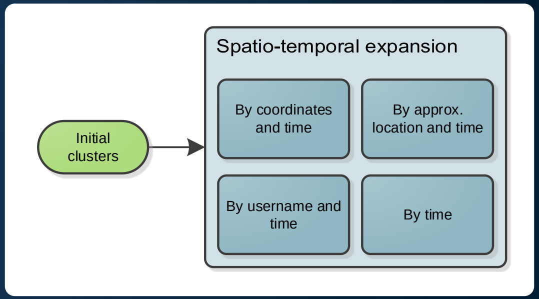
- Constraint: a person cannot be at multiple locations at the same time
- Relax constraint by linking it to a temporal duration for which it must hold
- For each user, determine location of location-unaware photos by majority voting w.r.t. location-aware photos that embed a similar capture date and time

Propagating location: *Approximate*

- Additionally: analyze each photo’s textual annotation (title, keywords, comments, …) for references to geographic locations
- *Countless* worldwide locations
  → limit search to larger cities
  → approximation
- Compile list of search locations using GeoNames dataset
- Use Linear Support Vector Classifier to limit search space
- Refine results based on text edit distances of consecutive work token combinations
- Associate *found* photos with geographic location
- Lastly: user-based propagation as before

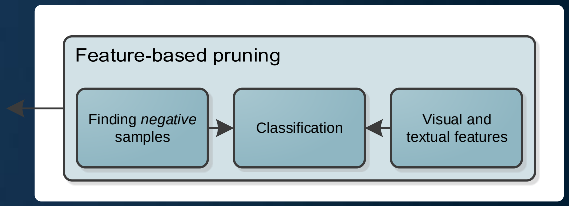
Spatio-temporal candidate expansion

- Start with exact/approximate spatio-temporal clusters
- Instead of limiting a retrieval space that may span an entire dataset → *expand* these spatio-temporal clusters by also including photos that do not embed location
- Expand based on: date and time, usernames, exact location, approximated location



Feature-based pruning

- Select or prune photos not belonging to retrieval space
- Train a binary model representing photos belonging or not belonging to a query (spatio-temporal event cluster)
- No separate training information available → compile a smaller random set of photos that do not intersect w.r.t. the date, time and location of a query
- Utilize a Linear Support Vector Classifier



Feature Extraction

Textual features:

- Utilize a roman pre-processor
- Apply a language-agnostic character-based tokenizer rather than a word-based tokenizer —> accommodate other languages as well as misspelled or varied terms
- Use TF-based vectorizer to convert tokens into a matrix of occurrences
- Limit amount of features to 9600 (*as good as* decomposition, but faster)

Visual features:

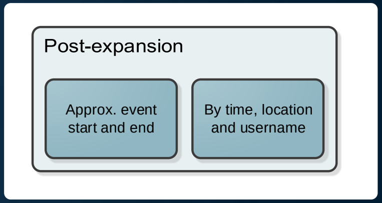
- GIST (a feature vector with 960 elements from a 4x4 image grid)

Feature fusion:

- Normalize both features
- Feature union (also incorporate weighting ratio)

Post-expansion

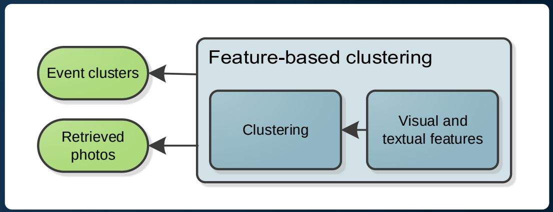
- Include photos that are likely relevant to the query but may have been *mistakenly* discarded by the prior feature-pruning step
- In particular, add photos that are linked to users who have multiple photos relevant to a query (event)
- Assumption: if a user attends a social event and takes photos, then it is likely that most of his photos taken over the time that he attends the event are *of* the event



Feature-based clustering

If a dataset includes mostly only photos according to events:

- K-Means clustering over entire dataset
- Same textual/visual features as in detection step
- Predict class labels of clusters by majority voting and by using the output of the event detection step



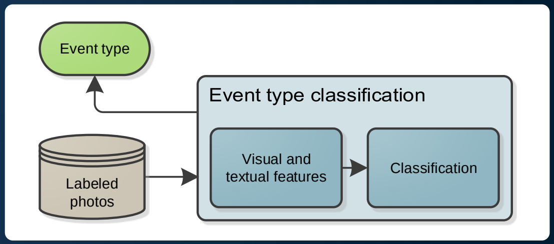
Event Type Classification

Basic approach:

- Expand ground truth that often only includes some photos of an event to multiple photos (using result of spatio-temporal clustering)
- Train a multi-class L-SVM based on textual/visual features
- Predict the event type of unknown *test* photos

Extension:

- Instead of treating unknown test photos separately, consider multiple test photos belonging to the same event together
- Perform majority vote: assign the most often predicted event type within an event to all its associated test photos



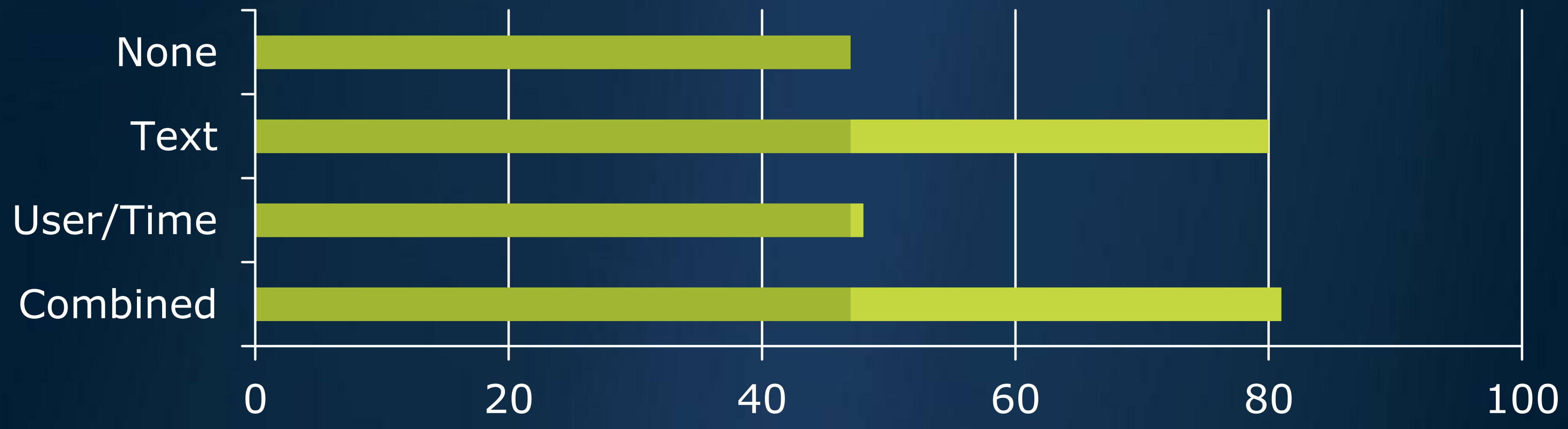
Experiments: Datasets

- 2013 MediaEval SED Dataset
- 306150 photos collected from Flickr (detection/retrieval)
- 57165 photos collected from Instagram (classification)
  → 9 event types (*sporting*, *protest*, *festival*, …, *other*)
- Metadata: unique photo ID, capture timestamp, username, title, description, keywords and partial geographic coordinates (partial, 46% and 27%)
- Ground truth in the form of event clusters with associated photos (specified by their photo IDs)
- Separate training set only for Instagram collection

Experiments: Datasets

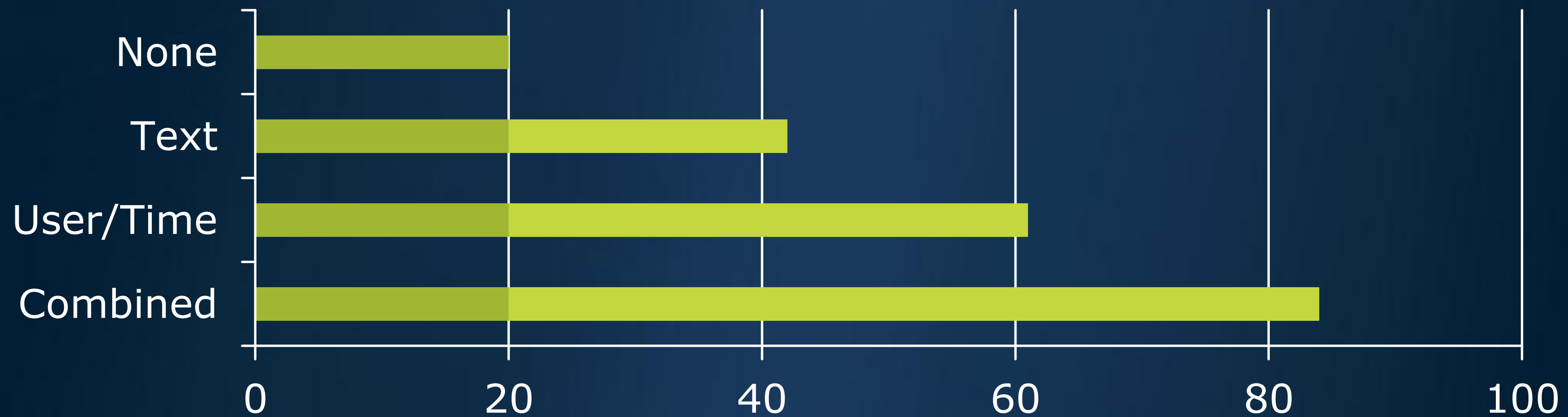
Results: spatio-temp. propagation

- 2013 SED dataset provides geographic coordinates for some (46%) but not all photos
- Able to approximate the location if based on textual information: by 33%
- Able to propagate and estimate the location if based on only the username: 1%
- Combined location propagation: by 34% to a total of 81%

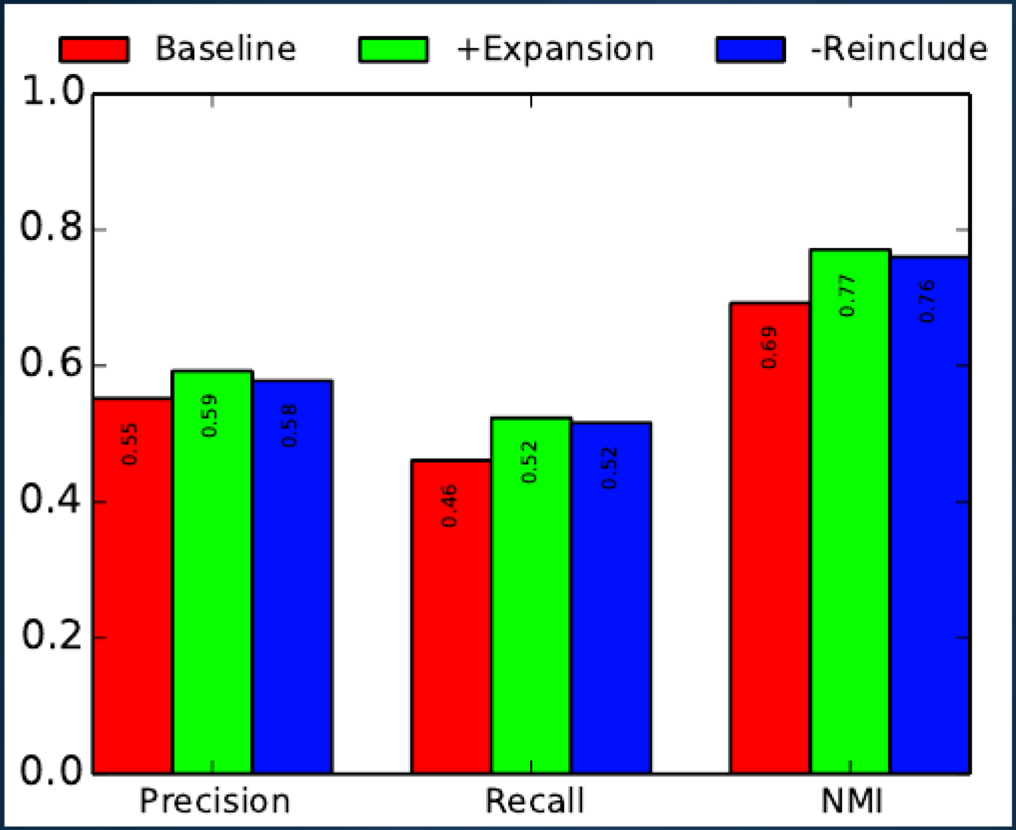


Results: spatio-temp. propagation

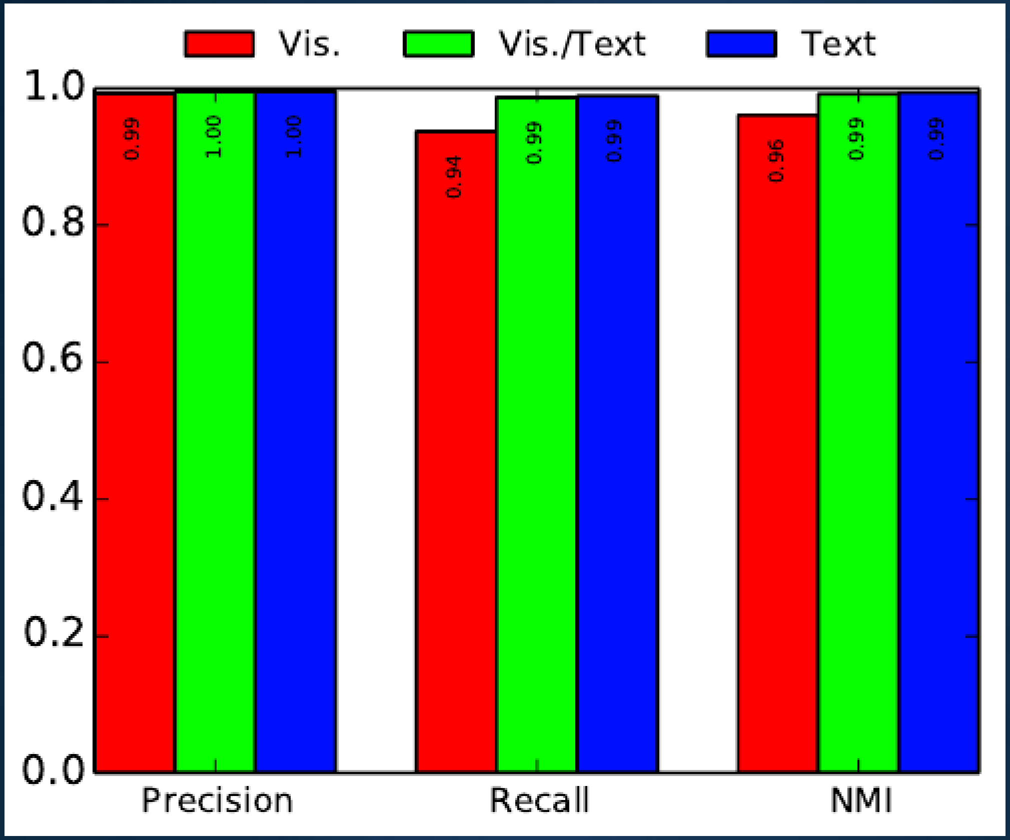
- 2012 SED dataset provides geographic coordinates for some (20%) but not all photos
- Able to approximate the location if based on only textual information: by 22%
- Able to propagate and estimate the location if based on only the username: by 41%
- Combined location propagation: by 64% to a total of 84%



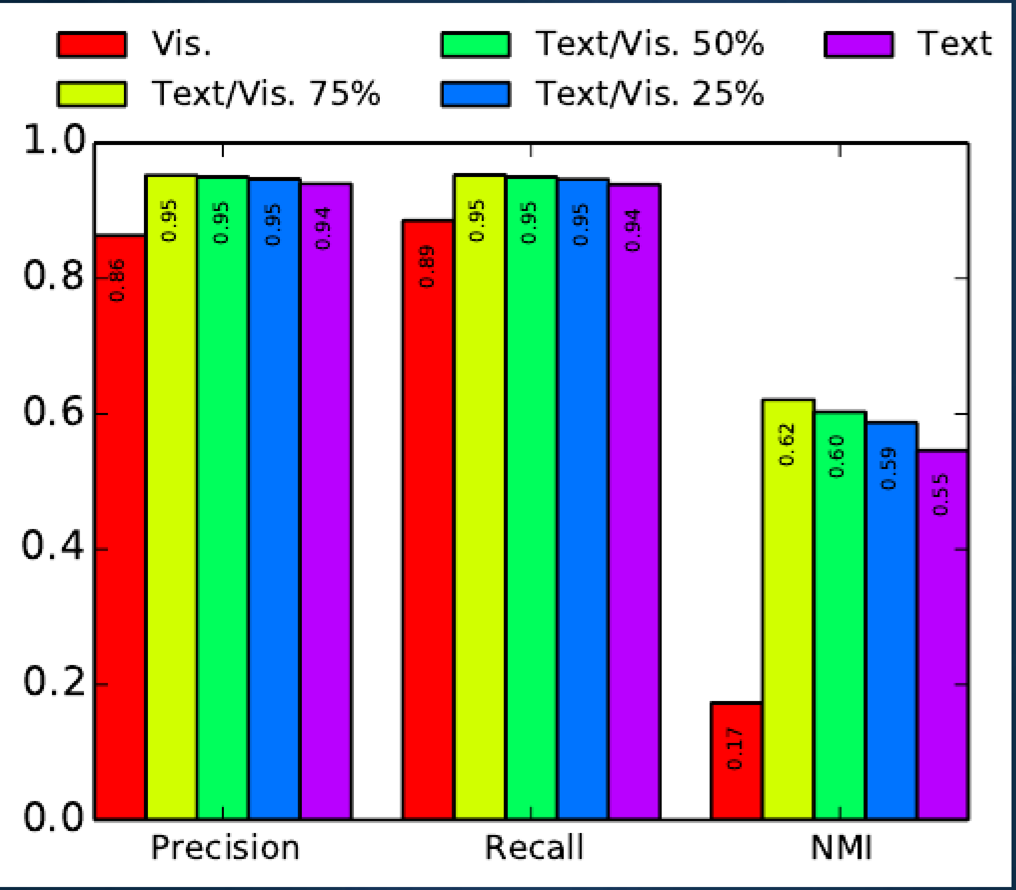
Results: basic detection/retrieval



Results: additional clustering



Results: event type classification



Results: event type classification

- Best performance for classifying as *non-event* instead of a particular event type
- Best performing types in terms of F1-score: *Concert* (0.52), *protest* (0.37), *theater-dance* (0.31)
- Worst performing types: *Fashion* and *other* (both under 0.1)

Conclusion

- Framework to retrieve photos associated with social events
- Operates on several domains (time, text, visual, etc.)
- Experiments suggest that:
  - Initial spatio-temporal propagation is vital to achieving good performance
  - Textual features notably outperform visual features
  - Additional clustering key for datasets that include only photo relating to events
- Future considerations: streaming operation, recurring events

Thank you!

Questions?