Lightweight Contextual Ranking of City Pictures: Urban Sociology to the Rescue

Vinicius Zambaldi †
fzvinicius@dcc.ufmg.br

Joao Paulo Pesce †
jpesce@dcc.ufmg.br

Daniele Quercia ‡
dquercia@acm.org

Virgilio Almeida †
virgilio@dcc.ufmg.br

† Universidade Federal de Minas Gerais, Brazil
‡ Yahoo Labs, Barcelona, Spain

Abstract

To increase mobile user engagement, photo sharing sites are trying to identify interesting and memorable pictures. Past proposals for identifying such pictures have relied on either metadata (e.g., likes) or visual features. In practice, techniques based on those two inputs do not always work: metadata is sparse (only few pictures have considerable number of likes), and extracting visual features is computationally expensive. In mobile solutions, geo-referenced content becomes increasingly important. The premise behind this work is that pictures of a neighborhood is linked to the way the neighborhood is perceived by people: whether it is, for instance, distinctive and beautiful or not. Since 1970s, urban theories proposed by Lynch, Milgram and Peterson aimed at systematically capturing the way people perceive neighborhoods. Here we tested whether those theories could be put to use for automatically identifying appealing city pictures. We did so by gathering geo-referenced Flickr pictures in the city of London; selecting six urban qualities associated with those urban theories; computing proxies for those qualities from online social media data; and ranking Flickr pictures based on those proxies. We find that our proposal enjoys three main desirable properties: it is *effective*, *scalable*, and aware of *contextual* changes such as time of day and weather condition. All this suggests new promising research directions for multi-modal learning approaches that automatically identify appealing city pictures.

1 Introduction

To offer an engaging mobile experience, photo sharing sites are trying to identify interesting and memorable geo-referenced pictures. To determine which pictures are interesting and memorable, researchers have heavily explored web-based solutions based on either metadata (e.g., likes) or visual features, or the combination of both. The main idea is that interesting pictures are those that have received a considerable number of likes or that contain the visual cues people often perceive to be beautiful.

Unfortunately, as we shall see in Section 2, metadata happens to be sparse (only few pictures have considerable number of likes), and visual extraction is computationally expensive and needs to be augmented with additional classes of features to guarantee good levels of accuracy.

To complement those solutions (largely meant for the web), we set out to consider a key element that has been hitherto left out: the idea of neighborhood. Pictures taken in a neighborhood reflect the neighborhood itself and people's idea of it. Since urban sociology has already dealt with those psychological aspects, we use prominent urban theories that aimed at explaining, for example, why a neighborhood is recognizable and distinctive (Lynch 1960; Milgram, Kessler, and McKenna 1972), and why it is perceived to be beautiful, quiet, and happy (Peterson 1967). In so doing, we make the following main contributions:

- We gather geo-referenced Flickr pictures and contextual variables (e.g., weather conditions) in the city of London (Section 3).
- We identify six main qualities that describe the way a city is psychologically perceived (Section 4) and quantify those qualities using proxies computed from Flickr and Foursquare data (Section 5).
- We rank Flickr pictures based on those proxies and find that such a ranking enjoys three main desirable properties (Section 6). First, it is *effective*, in that, the ranked results are similar yet complementary to the results produced by existing metadata-based solutions. Second, it is computationally inexpensive and, as such, *scalable*: our proxies are defined at the level of city rather than of picture and can be computed offline (no need for real-time updates). Third, it is aware of *contextual* factors (Section 7): different values of the same proxy can be computed as a function of, for example, the time of day or weather condition.

As we shall conclude in Section 8, these results suggest that, to offer a better mobile experience, future multi-modal learning research should further explore the idea of combining traditional features with domain-specific ones.

2 Related Work

To identify the pictures users tend to like, researchers have often used metadata. This is generally of two types. The first is *textual* metadata and is the most widely used: it consists of comments and tags users have annotated a picture with (van Zwol, Rae, and Garcia Pueyo 2010). The second type of metadata consists of *social* features and has received less

attention. van Zwol *et al.*, for example, used the communication and social network of Flickr users for predicting the number of likes (favorites) a picture has received (van Zwol, Rae, and Garcia Pueyo 2010). They found that social features alone yielded a good baseline performance, but the addition of textual features resulted in greatly improved precision and recall.

Despite showing good accuracies, approaches that rely on metadata suffer from coverage. That is because the frequency distributions of tags, comments, or any other social feature are power law: few pictures are heavily annotated, while many have little (if any) annotation (Sigurbjörnsson and van Zwol 2008). As such, approaches solely relying on metadata do not work for most of the pictures.

In those situations, researchers have explored the use of visual categorization. The most effective method is called bag-of-word model (Datta et al. 2006). This computes descriptors at specific points in an image. It has been shown that, given an image's descriptors, machine learning algorithms are able to predict whether people tend to find the image interesting and appealing (Redi and Merialdo 2012). The problem with visual categorization is that it is computationally expensive: it might take weeks to process 380 hours of video frames (van de Sande, Gevers, and Snoek 2011). To fix that, research effort has gone into designing faster methods and building new parallel computing architectures.

Within the multimedia research community, a considerable number of research papers have been proposing the *combined* use of metadata and visual features. These works employ multi-modal machine learning approaches that model topical, visual, and social signals together. Their goal has mainly been to predict which pictures users find appealing and aesthetically pleasing (van Zwol, Rae, and Garcia Pueyo 2010).

Those previous solutions have been designed to fit the general-purpose scenario of web ranking. However, when considering how pictures will be consumed on mobile phones, one might find that location becomes key: ranking pictures in location-based services might consider whether the neighborhoods in which the pictures were taken are highly visited, beautiful, or quiet. We set out to do just that by identifying desirable urban qualities from seminal work done in the 1970s.

3 Datasets

Within the bounding box of the city of London, we crawled 1.2M geo-referenced pictures using the Flickr public API. We also crawled their metadata, which includes: latitude and longitude points, number of comments, tags, upload date, taken date, number of favorites (those are Flickr's equivalent of likes), and number of views. The last two values have been used by past research as a signal of user preference for pictures (Yildirim and Süsstrunk 2013): the higher a picture's ratio of number of favorites to number of views, the more the picture's views have been converted into user likes. Figure 1 shows the density of photos in our dataset across London.

In addition to geo-referenced pictures, we collect data about two contextual factors. The first is 'time of the day'

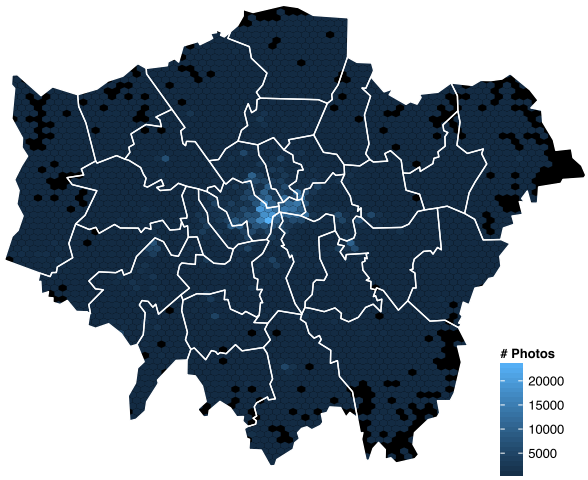

Figure 1: London Density Map of Photos in our Dataset.

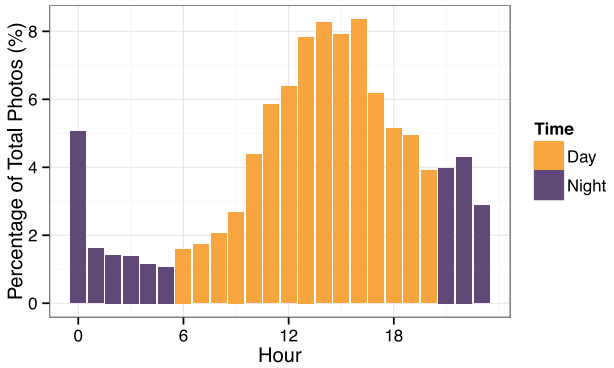

Figure 2: Fraction of Photos in each Hour of the Day ('day' is $[6am-10pm]$). We have 79.5% of the pictures being taken during the 'day' and 20.5% during 'night'.

and is computed based on the time each picture was taken: if it was taken between 6am and 10pm, we consider it to be taken during the 'day' (similar to what (Martínez and Santamaría 2012) did); otherwise, we consider it to be taken at 'night'. This results in 79.5% of the pictures being taken at 'day' and 20.5% at 'night' (Figure 2). Alternative temporal segmentations could have been chosen. We explored a variety of them and they all resulted in comparable percentages for day *vs.* night. The imbalance for number of pictures between day *vs.* night is natural as people tend to take more pictures during the day. However, this imbalance does not compromise any of our results as there are enough pictures at night to ensure statistically significance. The temporal span of the pictures in our dataset goes from 2002 to 2013.

The second contextual feature for which we collect data concerns weather conditions. We collect weather data from the British Atmospheric Data Centre for 11 years (2002-2013)[1]. This consists of hourly observation and amounts to roughly 10GB of data. We classify weather conditions as

[1]http://badc.nerc.ac.uk/data/ukmo-midas/WH_Table.html

follows: cloudy *vs.* not-cloudy; hot *vs.* cold; humid *vs.* dry; high visibility *vs.* low visibility; windy *vs.* not-windy.

4 Urban Qualities

Before mining those datasets, we need to identify the urban qualities that reflect people's psychological perceptions of the city.

4.1 Recognizability

Urban sociologists have suggested that the layout of urban spaces affects our sense of community well-being. Everyone living in an urban environment creates their own personal "mental map" of the city based on features such as the routes they use and the areas they visit. In his 1960 book "The Image of the City", Kevin Lynch hypothesized that the more recognizable the features of a city are, the more navigable the city is. Good imaginability allows city dwellers to feel at home (mental maps of good cities are economical of mental effort) and, as a result, their collective well-being thrives (Lynch 1960). To put his theory to test, researchers have recently used a web game that crowdsources Londoners' mental images of the city (Quercia et al. 2013). The researchers have replicated a well-known pen-and-paper experiment online: that experiment was run in 1972 by Milgram. He recruited his undergraduates in New York, showed them a variety of urban scenes, and asked them to guess the locations of those scenes (Milgram, Kessler, and McKenna 1972). Based on the correct answers, he drew the recognizability (collective mental) map his students had of New York. The web game replicates that experiment, in that, it picks up random urban scenes and tests users to see if they can judge the location in terms of closest subway station, borough, or region. In analyzing the results, the researchers found that areas suffering from social problems such as housing deprivation, poor living conditions, and crime are rarely present in residents' mental images. We use one of the researchers' aggregate datasets. This contains one recognizability score for each subway station, and another one for each borough in London. We have 150 tube stations and 30 boroughs before filtering, and 60 stations and 20 boroughs after filtering for unreliable scores.

4.2 Distinctiveness

It has been found that people recognize an area because of two main reasons: because they are exposed to it (e.g., a central area attracts residents from all over the city), and because the area offers a distinctive architecture (e.g., it hosts few star architects' buildings). In his 1972 article, Milgram simplified this idea by introducing the concept of distinctiveness. He stated that a place's recognizability can be expressed as $R_i = f(C_i \cdot D_i)$, where f is a function that predicts a place's recognizability from the centrality C_i of population flow (number of people who visit place i) and the place's social or architectural distinctiveness D_i. As a result, an area is socially or architecturally distinctive if its recognizability is not entirely explained by exposure to people but is also partly explained by its distinctiveness. In quantitative terms, this intuition translates into saying that i's distinctiveness D_i is the residual (error) of predicting R_i from C_i. As a proxy for flow C_i, we use the number of distinct subway passengers who have visited i.

Subway data. To compute the flow of subway passengers, we resort to an anonymized dataset containing a record of every journey taken on the London Underground using an Oyster card in the whole month of March 2010. Such cards are RFID-based technologies that replaced traditional paper-based magnetic stripe tickets in 2003. The dataset contains 76.6 million journeys made by 5.2 million users (each record consists of a passenger's trip from station a at time t_a, to station b at time t_b), and is available upon request from the public transportation authority.

Predicting distinctiveness. We have defined distinctiveness as the residual (error) of predicting an area's recognizability from its number of subway passengers. Thus, to compute distinctiveness, we need to predict recognizability first and then quantify the residual of such a prediction. Area i's *predicted* recognizability $\hat{R}_i$ is based on the number of unique subway passengers (denoted by C_i): $\hat{R}_i = \alpha + \beta {C_i}^2$. Such an expression assumes that the flow of subway passengers at a place is a good proxy for the number of people who have visited the place, which is likely the case in London given the widespread use of the underground (Smith, Quercia, and Capra 2013). Upon the predicted values $\hat{R}_i$, we can compute the area's distinctiveness D_i, which, given Milgram's original formulation, is the prediction error $D_i = R_i - \hat{R}_i$. In words, the less the flow of subway passengers predicts recognizability, the more distinctive the area is.

4.3 Eventfulness

In addition to studying whether an area is simply visited or not, one could also consider whether an area is routinely visited (e.g., the daily street from home to the train station) or whether it is visited in exceptional situations (e.g., during weekends or holidays). Previous work has partly shown that routine places are expected to be associated with georeferenced content that is less interesting than that associated with places that are visited in exceptional circumstances (Yildirim and Süsstrunk 2013).

To capture that intuition, we compute a measure that we call 'routine score'. We do so on a Foursquare dataset released by (Cheng et al. 2011): 22,387,930 Foursquare check-ins collected from September 2010 to January 2011. From these check-ins, we extracted those that happen to be in London: 230,785 check-ins in 8,197 places from 8,895 distinct users. To avoid computing anomalous scores, we filter out users with less than 10 check-ins and places which were visited by less than 10 distinct users. Then, for each user, we compute the fraction of times (s)he visits each location. By aggregating those user scores at each location (we used a geometric average as scores are skewed), we are able to

[2]The coefficients α and β are those for which the values of C_i are best predicted from the observed values R_i.

compute a location's routine score in the range $[0, 1]$: the higher it is, the more routine visits the location enjoys. To ease illustration, from the routine score, we derive its complementary measure, which we call 'eventfulness score' and is just 1 minus the routine score.

4.4 Beauty, Quiet, and Happiness

Not only mental maps but also aesthetically pleasing environments are associated with community well-being. Researchers in environmental aesthetics have widely studied the relationship between well-being and the ways urban dwellers perceive their surroundings (Nasar 1994; Taylor 2009; Weber, Schnier, and Jacobsen 2008). In 1967, Peterson proposed a methodology for quantifying people's perceptions of a neighborhood's visual appearance (Peterson 1967): he selected ten dimensions that reflected visual appearance (e.g., preference for the scene, greenery, open space, safety, beauty) and had 140 participants rate 23 pictures of urban scenes taken in Chicago along those dimensions (Peterson 1967). Based on his analysis, he concluded that preferences for urban scenes are best captured by asking questions concerning the beauty and safety of those scenes: beauty is synonymous with visual pleasure and appearance. To capture visual pleasure, the concept of *beauty* is thus key, and that is why it is our first perception quality. Beauty is indeed one of the three dimensions that recent work concerned with urban aesthetics has tried to quantify (Quercia, Ohare, and Cramer 2013). In this work, researchers collected votes on the extent to which hundreds of London urban scenes were perceived to be beautiful, quiet, and happy by more than 3.3K crowdsourcing participants. We get hold of the scores for beauty, quiet, and happiness at both subway and borough levels.

The researchers chose *quiet* because of popular discussions on 'city life'. Sound artist Jason Sweeney proposed a platform where people crowdsource and geo-locate quiet spaces, share them with their social networks, and take audio and visual snapshots. It is called Stereopublic[3] and is "an attempt to both promote 'sonic health' in our cities and offer a public guide for those who crave a retreat from crowds" - both for those in need of quietness and for people with disabilities, like autism and schizophrenia.

The remaining quality is that of *happiness*. This quality reflects the ultimate goal behind the 1970s research we have referred to: Milgram, Lynch and colleagues were after understanding which urban elements help to create intelligible spaces and would ultimately make residents happy.

Overall, we consider the three qualities of beauty, quiet, and happiness plus recognizability, distinctiveness, and eventfulness. Each of those qualities is defined at the two geographic levels of study: subway and borough levels.

5 Modeling Urban Qualities

To see how our urban qualities change depending on contextual factors, we need to build predictive models for each of them. To see why, consider our urban quality of beauty as an example. Its values could be represented on a heat map of London: darker squares (larger values) contain crowdsourced pictures considered to be beautiful, while lighter squares (smaller values) contain pictures considered to be less beautiful. One could then build a predictive model for beauty that estimates the extent to which those squares are dark (or light) on input of, say, Flickr or Foursquare metadata (e.g., likes on pictures, check-ins in Foursquare venues). By having this model at hand and stratifying the input metadata according to, say, time of day (e.g., number of favorites for photos taken at night), one could test which squares the model predicts to be beautiful at night, assuming that its predictions do not dramatically change with the contextual factors. We will test the validity of this assumption in Section 7.

The input features are derived from Flickr and Foursquare. These features include number of views, number of favorites, number of comments, number of tags, number of photos, number of unique Flickr users, number of unique Foursquare users, and number of check-ins. Since the urban qualities are defined at the levels of subway station and borough, we aggregate those features at the two levels. Then, if skewed, the features are log-transformed and, as such, their averages are not arithmetic but geometric.

On input of those features, we put the following models to test: linear model (least squares), decision tree regressor, support vector regression, ADA boost regressor, gradient boosting regressor, extra trees regressor and random forest regressor. For all the models, we have tried different parameter values and found that the default ones specified in the scikit-learn library[4] produced reasonable results. For brevity, we report only those results.

The predictive accuracies of the models are expressed with two measures: i) Mean Squared Error (MSE), which reflects the differences between the values predicted by the model under test and the actual values; and ii) Spearman's rank correlation ρ between two ordered lists of areas: in one list, areas are ranked by the model's predicted values; in the other list, areas are ranked by the actual values; ρ ranges from -1 to 1: it is 0 if the two lists are dissimilar, +1 if the two lists are exactly the same (best match), and -1 if the two lists are exactly reversed.

Figure 3 shows the models' error values (left panel) and accuracy values (right panel) for "in sample" predictions[5]. The large pink area reflects the statistical significance of the baseline being extremely low. The more sophisticated models (e.g., ADA boost, Gradient Boost) perform exceptionally well, yet simpler models (e.g., linear model, decision tree) show competitive performance: for all qualities other than quiet, the squared errors are below 0.03. The same goes for Spearman's ρ, which is always above 0.50 for all models. If we reduce the number of input features from 12 to 6, those results do not significantly change, suggesting that overfitting has little to do with such good prediction accuracies. To further reinforce this last point, we will now see to which extent such predicted values are associated with actual ap-

[3] http://www.stereopublic.net/

[4] http://scikit-learn.org/stable/

[5] We could not use cross-validation given the limited number of subway stations or boroughs.

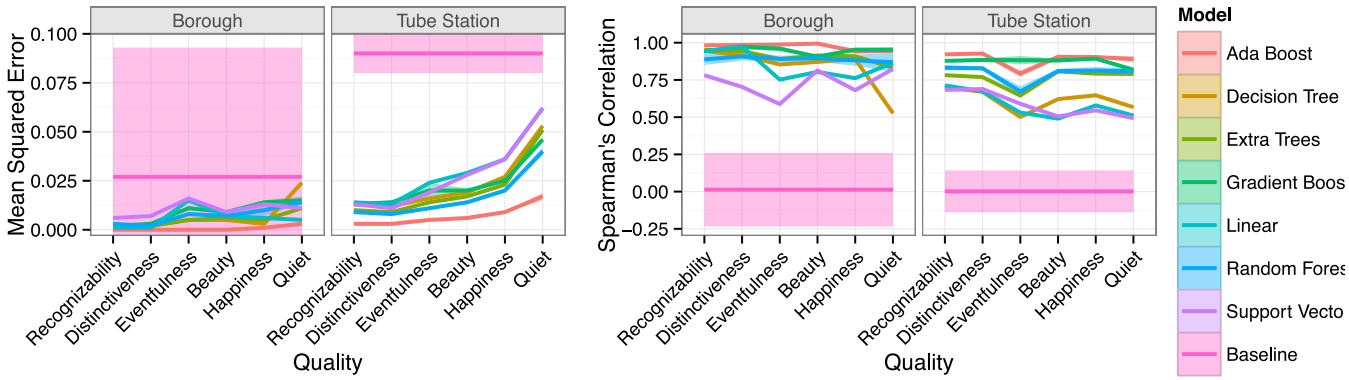

Figure 3: Mean Squared Error (left panel) and Spearman's Correlation ρ (right panel) for Area Rankings Produced by Seven Models plus Baseline. Each panel shows the results at both borough and subway station levels.

pealing content.

6 Rankings by Urban Qualities

We have just established how accurately off-the-shelf models can predict the urban qualities from Flickr and Foursquare metadata. However, we have not yet ascertained whether the predictions of those models would ultimately result into the selection of appealing geo-referenced pictures. To ascertain that, we need to determine which pictures are to be considered appealing. We do so by resorting to the widely-used normalized measure of community (user) appeal of picture i (Yildirim and Süsstrunk 2013):

$$\text{appeal}_i = \frac{\text{number of favorites}_i}{\text{number of views}_i}$$

The higher a picture's ratio of number of favorites to number of views, the more the picture's views have led to user likes. Pictures with few views do not need to be filtered away as their presence does not affect the overall ranking: pictures with many favorites and views will still be highly ranked.

We use the appeal measure to produce lists of geo-referenced pictures. Each list orders areas in a different way (we will see how) and, for each area, top k pictures ordered by appeal are, in turn, shown. Given that pictures are always ordered by appeal, the desirability of such a list depends on the ordering of areas. We produce two lists with two distinct orderings. In the first, areas are ordered at random (*baseline list*). In the second list, areas are ordered by a predicted urban quality (e.g., *beauty list*) [6]. As a result, both lists contain pictures that Flickr users have liked, but the order of areas in one list differs from that in other list. As such, by comparing the two lists, one can establish whether the urban qualities are useful for ranking city pictures or not. If there is no difference between the ways the two lists fare, then either the urban quality of, say, beauty does not happen to promote appealing geo-reference photos or its predicted values do not accurately reflect beautiful areas.

[6] We use an urban quality's predicted values and not the actual values to test to which extent our predictions are reasonable and whether they could be used in realistic scenarios.

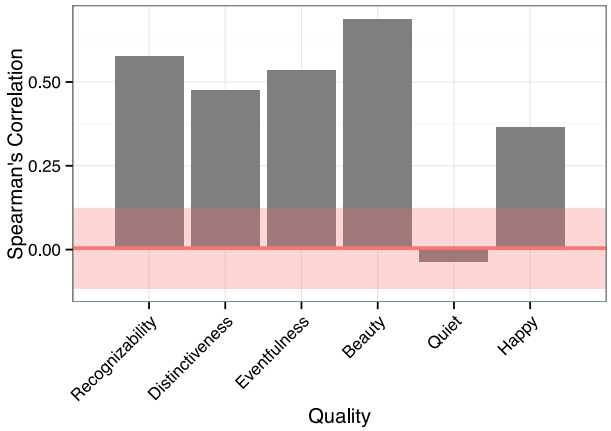

Figure 5: Similarity (Spearman's ρ) between the *Ideal List* and a List generated by one of our Urban Qualities. The similarity between baseline and the ideal list is shown in red with corresponding standard errors. For this barplot, the number of pictures per area is set to $k = 3$.

To quantitatively ascertain whether each of those two lists return appealing content, we build a third one, which we call *ideal list*: in it, pictures are ordered by appeal without any consideration for the areas in which they were taken. The more similar the *beauty list* to the *ideal list*, the more the urban quality of beauty is able to promote pictures that users have liked on Flickr. To measure the similarity of the two lists, we, again, use Spearman's rank correlation ρ.

Figure 5 shows the results, which suggest two noteworthy considerations. The first is that the *baseline list* greatly differs from the *ideal list* (as the red line shows) and differs from the remaining lists related to our urban qualities (suggesting that the ordering of *areas* matters). The second consideration is that the working hypothesis behind our work holds true: ordering areas by a given urban quality tends to preferentially promote city pictures that are indeed appealing. The quality that most successfully promotes appealing content is that of beauty ($\rho = 0.69$), followed by recognizability ($\rho = 0.58$), eventfulness ($\rho = 0.53$) and distinc-

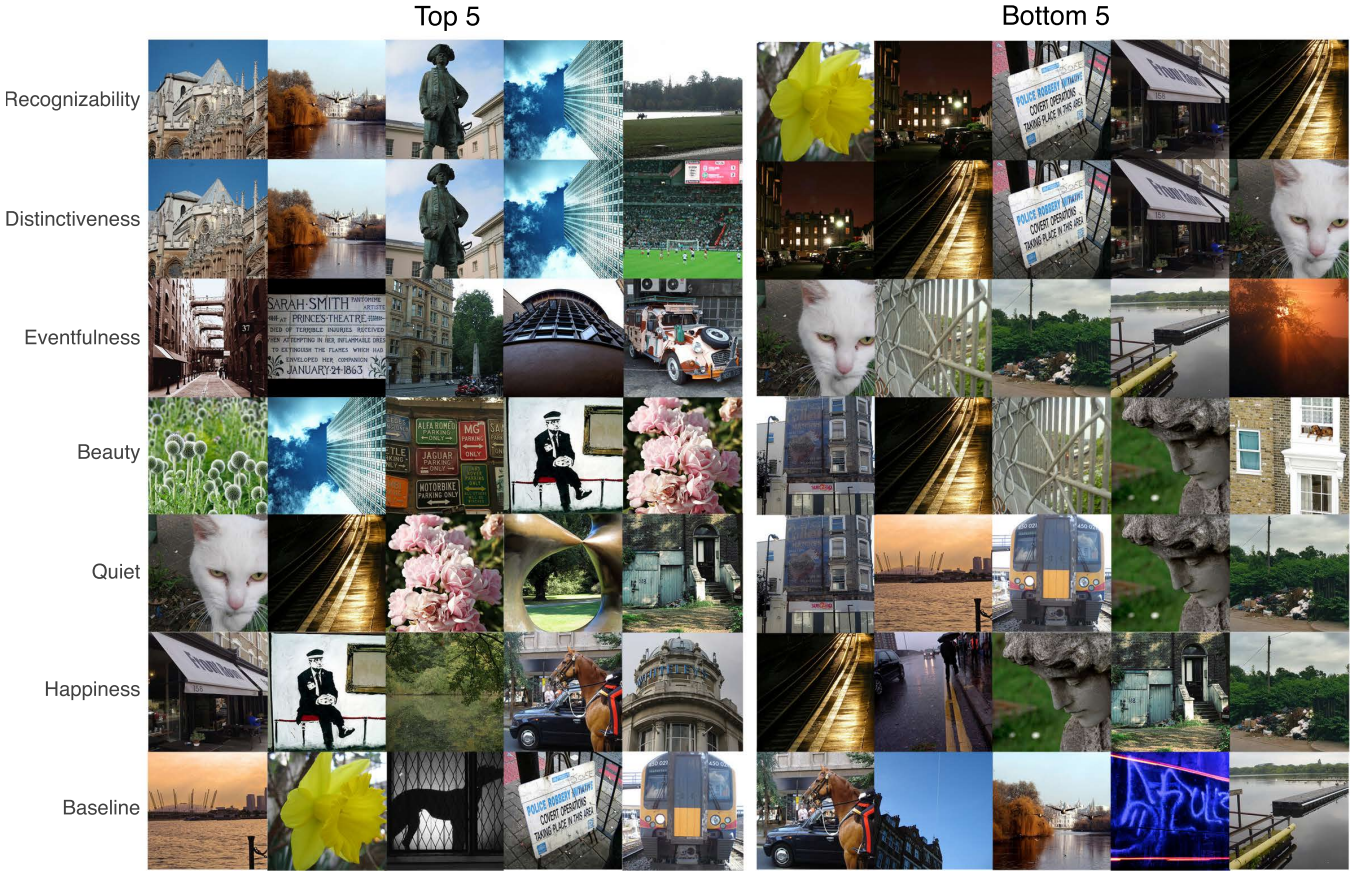

Figure 4: Pictures Ranked by Urban Qualities plus Baseline (last row). As an example, in the first row, top 5 (bottom 5) pictures in the five most (least), say, recognizable areas are shown.

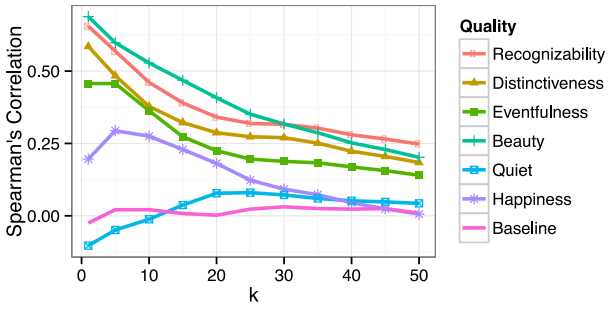

Figure 6: Similarity (Spearman's ρ) between the *Ideal List* and a List Generated by one of our Urban Qualities. The similarity varies with the number k of pictures per area (i.e., as the recommended list gets longer).

tiveness ($\rho = 0.47$). These results are further confirmed by visually inspecting the set of pictures ranked by each urban quality (Figure 4).

Figure 6 further shows that the Spearman correlation remains high as the user list of suggested pictures grows: suggesting five or even ten pictures in each area does not degrade the results at all. We also find that beautiful areas tend to be associated with appealing content, while quite areas are not (the rank by quiet is comparable to the baseline). This might be because quiet areas either are not associated with appealing content or are difficult to predict out of the metadata we have used here. Perhaps, further investigation should go into enlarging the pool of metadata to include textual descriptors or even city-wide sound recordings [7].

7 Contextual Factors

We now study how the predicted values of our urban qualities change depending on two contextual variables: time of day, and weather conditions.

To do so, in input of each of the models in the previous section, we give different features whose values change with the contextual variables. As we have mentioned in Section 5, this methodology is valid only if a model does not dramatically change with context. To test this assumption, we study whether the predictive accuracies of our models do not significantly change with time of day or weather, and we find this to be the case (Figure 7).

[7] http://cs.everyaware.eu/event/widenoise

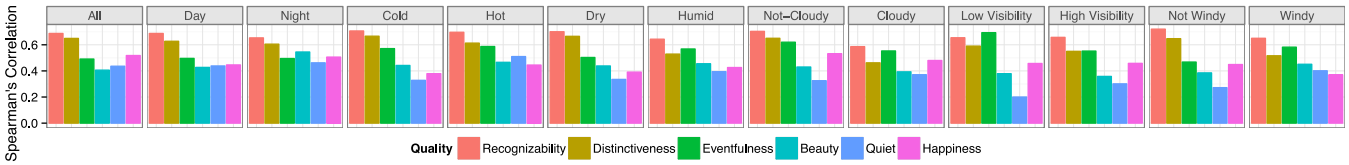

Figure 7: Accuracy of the Predicted Urban Qualities by Contextual Factors. Similarity (Spearman's ρ) between predicted and actual values for different contexts. The correlations do not significantly change for: day *vs.* night; cold *vs.* hot; dry *vs.* humid; not-cloudy *vs.* cloudy; low *vs.* high visibility; not-windy *vs.* windy.

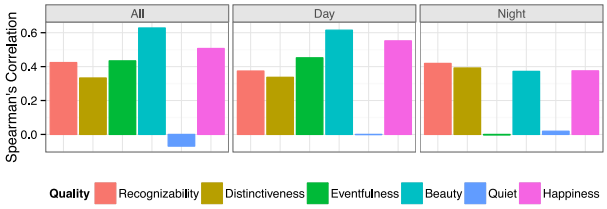

Figure 8: Rank Correlation (Spearman's ρ) Between the Ideal List and a List Generated by one of our Urban Qualities for Day *vs.* Night.

Figure 9: Pictures in Top 5 most Recognizable Areas at Day *vs.* Night.

Figure 11: Pictures in the Top 5 Most Recognizable Areas During Hot *vs.* Cold Days.

Weather Variable	Lower Condition	Upper Condition
Air temperature	cold	hot
Wet bulb temp.	dry	humid
Wind speed	not-windy	windy
Cloud level	not-cloudy	cloudy
Visibility	low-visibility	high-visibility

Table 1: Binary Discretization of Five Weather Variables.

.

7.1 Time of day

Using the definition of day *vs.* night in Section 3, Figure 8 shows the similarity (Spearman ρ) between the ideal list and a list generated by a given urban quality during different times of the day. The higher the similarity, the more the generated list contains appealing content. We find that beautiful areas tend to be associated with appealing content more during the day than during the night (the cerulean bar decreases from day to night). In a similar way, eventful areas are associated with appealing content during the day, which might reasonably suggest that people do not tend explore new parts of the city at night. Also, by visually inspecting the pictures in the top 5 most recognizable areas at day *vs.* those in the top 5 at night (Figure 9), one observes two distinctive sets of results, which speaks to the external validity of our approach.

7.2 Weather

For every day present in our weather dataset between 2002 and 2013, we discretize each of the five weather variables listed in Table 1 into lower class and upper class depending on whether their values are in the bottom or upper quartiles (Table 2 shows the resulting thresholds). Depending on the weather condition of the day a picture was taken, we associate the five discretized values with the picture. For example, for a photo taken at 2007-06-09 17:05, its associated weather variables are: wind speed is $2knots$, air temperature is $24.7°C$, wet bulb temperature is $18.0°C$, cloud level is $6oktas$, and visibility is $12km$. That translates into associating the following discretized values with the picture: hot, humid, not-windy, low-visibility, and not-cloudy. Table 3 shows the fraction of photos taken under different weather conditions: as one expects, photos are taken in non-cold and non-dry days; also, people tend to avoid cloudy days while preferring high visibility days.

Figure 10 shows the similarity (Spearman ρ) between the ideal list and a list generated by a given urban quality under different weather conditions. We find that, with hot weather (which, in London, means a temperature above 16 degrees Celsius), any type of area (whether it is recognizable, distinctive, eventful, beautiful, or happy) is associated with appealing content. Dry and cold turn out to be the weather conditions that most negatively affect the production of appealing content. Again, ranking pictures during hot *vs.* cold days results in meaningful and inexpensive segmentations (Figure 11).

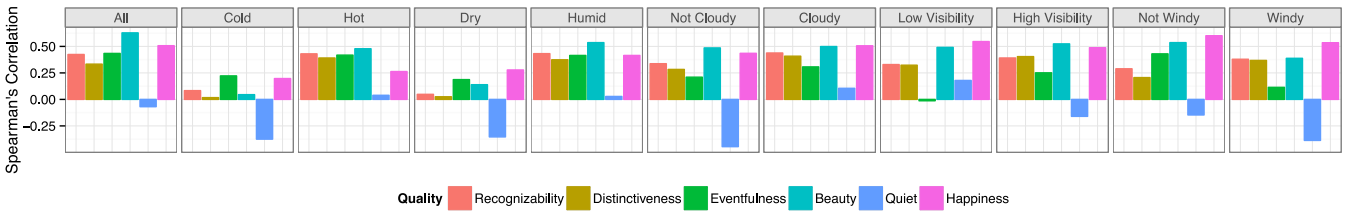

Figure 10: Rank Correlation (Spearman's ρ) Between the Ideal List and the Lists Generated by one of our Urban Qualities across Different Weather Conditions.

Weather Condition	t_{lower}	t_{upper}	Units
Cold/Hot	7.2	15.9	$degC$
Dry/Humid	5.8	13.2	$degC$
Not-windy/Windy	5.0	11.0	$knots$
Not-cloudy/Cloudy	2.0	8.0	$oktas$
Low-visible/High	12.0	29.0	km

Table 2: Upper and Lower Thresholds (t_{lower} and t_{upper}) used to Discretize Weather Variables.

.

Weather Condition	$\% < t_{lower}$	$\% > t_{upper}$	outside
Cold/Hot	16.4	40.5	43.1
Dry/Humid	17.5	36.0	46.5
Not-windy/Windy	24.2	31.8	44.0
Not-cloudy/Cloudy	34.4	26.8	38.7
Low-visible/High	18.9	27.3	53.8

Table 3: Fraction of Photos Under Different Weather Conditions.

8 Discussion

We now discuss the main limitations of this work, and how to frame it within the context of emerging research.

Limitations. This work is the first step towards using urban features to identify appealing geo-referenced content. In the future, research should go into combining all classes of features together. One simple way of doing so is to order each area's pictures depending on how appealing they are (appeal can be derived from visual features). The second limitation is that new ways of presenting pictures other than segmenting them by city neighborhoods (which are politically-defined and might be arbitrary at times) are in order: one could, for example, show pictures by areas that emerge from location-based data. Cranshaw *et al.* (Cranshaw et al. 2012) used Foursquare data to draw dynamic boundaries in the city: what they called 'livehoods'. However, any work that uses location-based data (including ours) should account for the limitation of the data itself: the geographic distribution of Foursquare check-ins is biased (Rost et al. 2013) (e.g., a user is likely to check-in more at restaurants than at home), and that can greatly affect the computation of our routine scores. Finally, given our promising results, it might be beneficial to further explore the use of urban features in cold-start situations, which are increasingly common.

Complementary to existing approaches. This work has to be considered complementary to existing approaches. By no means, it is meant to replace ranking solutions based on metadata or on visual features. Instead, all these solutions can be used together considering that they work under different conditions: whenever pictures come with rich metadata, then that metadata could be used to rank them; by contrast, in cold-start situations, our lightweight ranking combined with visual features might well be the only option at hand. We have shown that this option is viable as it offers good baseline performance. More generally, our results speak to the importance of incorporating cross-disciplinary findings. This work heavily borrows from 1970s urban studies and is best placed within an emerging area of Computer Science research, which is often called 'urban informatics'. Researchers in this area have been studying large-scale urban dynamics (Crandall et al. 2009; Cranshaw et al. 2012; Noulas et al. 2012), and people's behavior when using location-based services such as Foursquare (Bentley et al. 2012; Cramer, Rost, and Holmquist 2011; Lindqvist et al. 2011).

9 Conclusion

In the web context, the problem of automatic identification of appealing pictures has been often casted as a ranking problem. By contrast, in the mobile context, we posited that the research roadmap should differ and revolve around the concept of neighborhood. Before this work, we did not know whether and, if so, how some of the 1970s theories in urban sociology could be practically used to identify appealing city pictures. We have shown that, upon theories proposed by Lynch, Milgram and Peterson, one is indeed able to do so. We hope that these results will encourage further work on multi-modal machine learning approaches that combine traditional (e.g., visual, textual, and social) features with domain-specific urban features.

References

Bentley, F.; Cramer, H.; Hamilton, W.; and Basapur, S. 2012. Drawing the city: differing perceptions of the urban environment. In *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI)*.

Cheng, Z.; Caverlee, J.; Lee, K.; and Sui, D. Z. 2011. Ex-

ploring Millions of Footprints in Location Sharing Services. In *ICWSM*.

Cramer, H.; Rost, M.; and Holmquist, L. E. 2011. Performing a check-in: emerging practices, norms and 'conflicts' in location-sharing using foursquare. In *Proceedings of ACM International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI)*.

Crandall, D. J.; Backstrom, L.; Huttenlocher, D.; and Kleinberg, J. 2009. Mapping the world's photos. In *Proceedings of ACM International Conference on World Wide Web (WWW)* .

Cranshaw, J.; Schwartz, R.; Hong, J.; and Sadeh, N. 2012. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Datta, R.; Joshi, D.; Li, J.; and Wang, J. Z. 2006. Studying Aesthetics in Photographic Images Using a Computational Approach. In *Proceedings of the 9th European Conference on Computer Vision (ECCV)*.

Lindqvist, J.; Cranshaw, J.; Wiese, J.; Hong, J.; and Zimmerman, J. 2011. I'm the mayor of my house: examining why people use foursquare - a social-driven location sharing application. In *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI)*.

Lynch, K. 1960. *The Image of the City*. Urban Studies. MIT Press.

Martínez, P., and Santamaría, M. 2012. Atnight: Visions through data. *Mass Context* 172–183.

Milgram, S.; Kessler, S.; and McKenna, W. 1972. A Psychological Map of New York City. *American Scientist*.

Nasar, J. L. 1994. Urban design aesthetics: The evaluative qualities of building exteriors. *Environment and Behavior* 26(3):377.

Noulas, A.; Scellato, S.; Lambiotte, R.; Pontil, M.; and Mascolo, C. 2012. A Tale of Many Cities: Universal Patterns in Human Urban Mobility. *PLoS ONE*.

Peterson, G. L. 1967. A Model of Preference: Quantitative Analysis of the Perception of the Visual Appearance of Residential Neighborhoods. *Journal of Regional Science* 7(1):19–31.

Quercia, D.; Pesce, J. P.; Almeida, V.; and Crowcroft, J. 2013. Psychological Maps 2.0: A web gamification enterprise starting in London. In *Proceedings of ACM International Conference on World Wide Web (WWW)* .

Quercia, D.; Ohare, N.; and Cramer, H. 2013. Aesthetic Capital: What Makes London Look Beautiful, Quiet, and Happy?

Redi, M., and Merialdo, B. 2012. Where is the Beauty?: Retrieving Appealing Video Scenes by Learning Flickr-based Graded Judgments. In *Proceedings of the 20th ACM Conference on Multimedia (MM)*.

Rost, M.; Barkhuus, L.; Cramer, H.; and Brown, B. 2013. Representation and communication: Challenges in interpreting large social media datasets. In *16th ACM Conference on Computer Supported Cooperative Work and Social Computing*.

Sigurbjörnsson, B., and van Zwol, R. 2008. Flickr Tag Recommendation Based on Collective Knowledge. In *Proceedings of the 17th ACM Conference on World Wide Web (WWW)*.

Smith, C.; Quercia, D.; and Capra, L. 2013. Finger On The Pulse: Identifying Deprivation Using Transit Flow Analysis. In *Proceedings of ACM International Conference on Computer-Supported Cooperative Work (CSCW)* .

Taylor, N. 2009. Legibility and aesthetics in urban design. *Journal of Urban Design* 14(2):189–202.

van de Sande, K. E.; Gevers, T.; and Snoek, C. G. 2011. Empowering Visual Categorization With the GPU. *IEEE Transactions on Multimedia* 13(1).

van Zwol, R.; Rae, A.; and Garcia Pueyo, L. 2010. Prediction of Favourite Photos Using Social, Visual, and Textual Signals. In *Proceedings of ACM Conference on Multimedia (MM)* .

Weber, R.; Schnier, J.; and Jacobsen, T. 2008. Aesthetics of streetscapes: Influence of fundamental properties on aesthetic judgments of urban space 1, 2. *Perceptual and motor skills* 106(1):128–146.

Yildirim, G., and Süsstrunk, S. 2013. Rare is Interesting: Connecting Spatio-temporal Behavior Patterns with Subjective Image Appeal. In *Proceedings of the 2nd ACM International Workshop on Geotagging and Its Applications in Multimedia (GeoMM)*.